Eigennamenerkennung in Web-Korpora des Deutschen. Eine Herausforderung für die (Computer)linguistik

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Arts (B.A.)
im Fach Germanistische Linguistik

Humboldt-Universität zu Berlin
Philosophische Fakultät II
Institut für deutsche Sprache und Linguistik

eingereicht von Lea Arianna Helmers

Erstgutachterin: Prof. Dr. Anke Lüdeling
Zweitgutachter: Dr. Roland Schäfer

Berlin, den 25. Juli 2013

Inhaltsverzeichnis

1 Zu Eigennamen und ihrer Relevanz für die linguistische Forschung

Die Entscheidung, den Beitrag über einen Dackel *Herzkranke Emmy nach 26 Stunden aus Erdloch befreit*[1] zu betiteln und anstelle der appellativischen Gattungsbezeichnung den Eigennamen (EN) der Hündin zu verwenden, war vermutlich eine bewusste. Namen haben einen besonderen Stellenwert in unserer Gesellschaft. Dies lässt sich nicht zuletzt daran festmachen, dass Menschen sie hauptsächlich an Dinge und Lebewesen vergeben, mit denen sie emotional verbunden sind oder denen sie Individualität zuerkennen wollen. Man vergibt sie an Kinder und Haustiere, aber auch Waren werden beispielsweise mit Namen versehen, um sie als besonders und wertvoll darzustellen. Im Falle der Bild-Schlagzeile führt die Verwendung des EN *Emmy* somit zu einer Individualisierung der betroffenen Hündin, wodurch die Emotionen des Lesers verstärkt werden sollen. Ähnliche Beispiele lassen sich nicht nur in Boulevardblättern finden und stellen ein gängiges Mittel dar, um Aufmerksamkeit zu erregen. Dass EN einen derart besonderen Status in unserer Gesellschaft haben, macht verständlich, dass sich viele verschiedene Disziplinen mit ihrer Erforschung beschäftigen. So interessieren sich nicht nur Linguisten für Onyme, sondern unter anderem auch Psychologen, Soziologen und Marketingforscher. Nicht zuletzt sind sie auch ein wichtiger Anhaltspunkt für maschinelle Sprachverarbeitungsprozesse wie Informationsextraktion oder die automatische Beantwortung von Fragen. Für akkurat funktionierende Systeme, die solche Anwendungen ausführen, ist die Markierung von EN in dem zu verarbeitenden Text eine wichtige Voraussetzung. In dieser Arbeit sollen EN jedoch vorrangig unter dem Gesichtspunkt der Onomastik betrachtet werden, deren Aufgabe es ist, die linguistischen Eigenschaften von EN zu beschreiben und zu systematisieren. Hierfür sind, wie in allen linguistischen Gebieten, Datengrundlagen notwendig, anhand derer die Besonderheiten der EN untersucht werden können. Korpora scheinen hierfür sehr geeignet, da sie eine leicht zugängliche Quelle für EN-Belege sind, anhand derer auch das syntaktische Umfeld der Onyme untersucht werden kann. Für eine effiziente Arbeit mit solchen Korpora ist jedoch, vor allem, wenn es sich bei letzteren um umfangreiche Textsammlungen handelt, eine vorangehende Aufbereitung, welche die Kennzeichnung von EN mit einschließt, unabdingbar. Bisher haben automatische Tagger allerdings selbst auf Zeitungsdaten für das Deutsche kaum zufriedenstellende Ergebnisse bei der EN-Erkennung erreicht. Gerade in großen Korpora stellt eine manuelle Aufbereitung nichtsdestotrotz keine Alternative zur maschinellen Aufbereitung dar, weshalb dringender Bedarf besteht, die automatische EN-Kennzeichnung zu verbessern.

Mit der vorliegenden Arbeit möchte ich hierzu insofern einen Beitrag leisten, als ich in ihrem Rahmen die zur Zeit frei verfügbare Software für die EN-Erkennung in deutschen Texten evaluiert habe. Hierbei wurden allerdings keine Zeitungsdaten oder ähnlich editierte Texte als Testmaterial gewählt, die zweifelsohne eine große

[1] BILD BERLIN, 25.04.2013, S. 3.

Homogenität mit den von den Entwicklern verwendeten Trainings- und Testkorpora aufweisen. Stattdessen wurden die Programme meines Wissens erstmalig auf einem Web-Korpus getestet, welches mit Material aus Foren und ähnlichen Kommunikationsplattformen auch kaum editierte Texte enthält. Diese Datengrundlage erlaubt es, einen realistischeren Eindruck von der Performanz der EN-Erkenner zu erhalten, da eher spontansprachliches Material einen beachtlichen Teil unseres Sprachgebrauchs ausmacht, es aber in den meist stark redigierten Zeitungen kaum vorkommt. Ähnlich wie **giesbrecht09** es für Part-of-Speech-Tagger (POS-Tagger) zeigten, ist zu erwarten, dass diese Daten sich als besondere Herausforderung für die EN-Erkenner erweisen. Da das Internet allerdings eine höchst interessante, leicht zugängliche und umfangreiche Sammlung natürlicher Sprachdaten darstellt, ist zu vermuten, dass die Bedeutung von Web-Korpora als Grundlage für die Bearbeitung von Forschungsfragen in Zukunft zunehmen wird, weshalb Sprachverarbeitungstools auch auf solchen Daten zuverlässige Ergebnisse liefern sollten.
Aufgrund des Umstands, dass eine Auswertung der Performanz lediglich mögliche Unzufriedenheiten mit der Güte der Programme aufzeigt, jedoch kaum konstruktive Vorschläge für eine Verbesserung der Software direkt daraus abgeleitet werden können, schließt sich an die Evaluation eine detaillierte Fehlerklassifikation an. Anhand der Ergebnisse derselben und dem Vergleich der Programme werden daraufhin die unterschiedlichen Lernansätze, die die Entwickler für ihre Programme verwendeten, besprochen und mögliche Maßnahmen für die Verbesserung der Performanz erwogen.
Der Aufbau der Arbeit sieht vor, dass zunächst in einem einführenden Kapitel grundlegende funktionale und grammatische Eigenschaften von EN in Anlehnung an die aktuelle und umfassende Einführung in die Onomastik von **nuebling12** dargestellt werden, da diese sowohl für die linguistische Definition als auch für die Erkennung der EN eine wichtige Rolle spielen. Im darauf folgenden Kapitel werden dann die im Rahmen dieser Arbeit evaluierten EN-Erkenner vorgestellt und die Performanz auf den ursprünglichen Testdaten der Entwickler wiedergegeben. Kapitel vier enthält neben den Evaluationsergebnissen eine genauere Beschreibung des hier als Testkorpus verwendeten CatTle.de.12 und der Richtlinien, auf deren Grundlage die Programme evaluiert wurden. Im fünften Kapitel wird eine detaillierte Fehlerklassifikation vorgenommen und mithilfe eines statistischen Verfahrens verschiedene Eigenschaften der Tokens auf ihren Einfluss auf die Fehlerentstehung überprüft. Auf dieser Grundlage können dann im sechsten Kapitel das Potential und die Grenzen der unterschiedlichen für die EN-Erkennung relevanten Lernmechanismen sowie Ansätze zur Verbesserung der Performanz beleuchtet werden.

2 Linguistische Charakteristiken der Eigennamen

Die Annotation von EN in einem Text setzt zunächst Kriterien voraus, nach denen entschieden werden kann, ob es sich bei einem Token um ein Onym, bzw. einen

Teil eines solchen handelt oder nicht. In vielen Fällen helfen besondere Eigenschaften von EN durchaus dabei, sie von anderen Nomen und insbesondere von den ihnen nahestehenden Appellativen, also den Gattungs- und Klassenbezeichnungen, abzugrenzen. Zum Beispiel handelt es sich bei ihnen nicht selten um Ausdrücke, die zunächst nicht im eigentlichen Wortschatz der Sprache enthalten sind. Damit geht unter anderem einher, dass sie auf vielen linguistischen Ebenen von anderen Substantiven abweichen. Allerdings ist dies nicht immer der Fall und es gibt eine Vielzahl von Appellativen, die auch onymisch verwendet werden können. Die in **nuebling12** besprochenen Uneinigkeiten, die über die Einteilung in Appellative und EN unter Namensforschern herrschen, zeigen, dass sich in einigen Fällen die eindeutige Zuordnung eines Wortes zu einer der beiden Kategorien als schwierig oder gar unmöglich erweist. Vielmehr ist die Klassifizierung also von Grenz- und Zweifelsfällen gezeichnet, was nicht zuletzt auch dem Umstand geschuldet ist, dass sich ein Großteil der EN aus Appellativen entwickelt hat (vgl. **nuebling12**). Ist ein Phänomen theoretisch schon nicht klar umrissen und sind die Kategorien nicht eindeutig abgrenzbar, so ist zu erwarten, dass sich diese Unsicherheiten auch in der automatischen Annotation widerspiegeln. Dennoch sind für eine Auswertung von Programmen, die Eigennamen erkennen, klare Entscheidungen und deren konsequente Einhaltung unabdingbar. Aus diesem Grund verfolge ich in dieser Arbeit den Ansatz einer prototypischen Klassifizierung, wobei die in **nuebling12** genannten funktionellen und grammatischen Unterschiede zwischen Appellativen und EN für die von mir getroffenen Entscheidungen ausschlaggebend waren. Auf diese möchte ich nun im Folgenden genauer eingehen und dabei auch auf ihr Potential als Merkmale für eine korrekte automatische Klassifikation hinweisen.

2.1 Funktion von Eigennamen und Appellativen

Ein grundsätzlicher Unterschied von Appellativen und EN besteht in ihrer Referenzleistung. Während appellativische Bezeichnungen stets eine Charakterisierung ihres Denotats und damit eine lexikalische Semantik beinhalten, erfolgt bei EN die Referenz auf das Bezeichnete ohne die „Aktivierung einer potentiellen Bedeutung“ (**nuebling12**). Das heißt, die Referenz von EN auf ihr Signifikat geschieht ohne die Zwischenebene eines lexikalischen Inhalts, weshalb man von Direktreferenz spricht. Des Weiteren bezeichnen EN im Idealfall nur ein Denotat und werden daher monoreferent genannt. Appellative hingegen referieren auf Klassen von Objekten, die einige, aber in der Regel nicht alle Merkmale gemeinsam haben. Die Eigenschaft der Monoreferenz erklärt auch das individualisierende und identifizierende Potential von EN, welches bereits in Bezug auf die Verwendung des EN in der Bild-Schlagzeile erwähnt wurde, und hilft bei der Kategorisierung von einigen Zweifelsfällen. So wird unter Berücksichtigung des Monoreferenzkriteriums schnell klar, dass Monats- und Wochentagsbezeichnungen zu den Appellativen zählen, es jedoch durchaus Zeitnamen (Chromonyme) wie *das Mittelalter* oder *das Holozän* gibt, die auf einen abge-

schlossenen, einzigartigen Zeitraum referieren.
Dass die Eigenschaft eines Substantivs, monoreferent zu sein, jedoch keine hinreichende Bedingung für dessen Kategorisierung als EN darstellt, machen die sogenannten Unika wie *Sonne*, *Mond* oder *Paradies* deutlich, die außerhalb der physikalischen Fachlexik nur ein Denotat haben, welches sie jedoch durch ihren semantischen Gehalt charakterisieren. **nuebling12** zählen sie deshalb zu den Appellativen. Ebenso wenig handelt es sich bei der Monoreferenz um eine notwendige Bedingung für die Kategorie EN insofern, als auch Warennamen wie *Snickers* oder *Berliner Pilsner* als EN kategorisiert werden. Wenngleich diese meist eine große Anzahl von Objekten bezeichnen und Monoreferenz folglich nicht gegeben ist, sind sie dennoch direktreferent und werden ihrem Denotat in einem Referenzfixierungssakt zugeordnet.
Ebenso gibt es Onyme, bei denen Direktreferenz nicht im eigentlichen Sinne vorliegt. Dies trifft vor allem auf die sogenannten Gattungs-EN wie *Bodensee*, *Wallensteingraben* oder *Schwarzes Meer* zu, welche die appellativische Bezeichnung ihres Denotats beinhalten. Oft ist hier der Übergang zwischen Appellativ und EN fließend, sodass es nicht immer leicht ist, zu entscheiden, welcher Kategorie ein solches Substantiv angehört. Folglich lässt sich vermuten, dass derartige Fälle auch für automatische EN-Erkenner ein Problem darstellen.

2.2 Grammatische Unterschiede von Appellativen und Eigennamen

Über die Referenzleistung hinaus unterscheiden sich EN und Appellative in mehreren ihrer linguistischen Eigenschaften. Dies betrifft vor allem die Ebene der Phonologie und Prosodie und jene der Morphologie, der Morphosyntax und der Graphematik. Da Prosodie sich in der Regel nicht im Schriftbild widerspiegelt und somit für die automatische EN-Erkennung keine Rolle spielt, möchte ich hier ebenso wie auf die nach **nuebling12** wenigen syntaktischen Unterschiede nicht näher eingehen, sondern mich auf die übrigen Beschreibungsebenen konzentrieren.

2.2.1 Phonologische Besonderheiten

EN können phonologisch stark vom üblichen Wortschatz abweichen. So enthalten sie häufig Konsonanten- oder Vokalkombinationen, die in der betreffenden Sprache sonst nicht vorkommen, wie beispielsweise die im deutschen Lautsystem sonst nicht vorhandene Phonem-Abfolge /mʀ/ im Familiennamen *Mrotzek* oder der Diphthong /ua/ in *Eduard*. Zudem können in EN auch Vollvokale in Nebensilben, die im Deutschen sonst ausschließlich den Schwa-Laut enthalten, auftreten. Unter der Berücksichtigung des Umstands, dass EN oft aus anderen Sprachen entlehnt sind, ist dies nicht weiter verwunderlich. Da sich diese besonderen Lautkombinationen auch schriftlich manifestieren, könnten sie unter Umständen bei der automatischen EN-Erkennung von Nutzen sein.

2.2.2 Morphologische Besonderheiten

Auch in ihren morphologischen Eigenschaften unterscheiden sich EN von anderen Substantiven. Sowohl was die Kasusmarkierung als auch die Pluralbildung betrifft, weisen Onyme eine Minimalflexion mit Tendenz zur Deflexion auf, wobei letztere im Akkusativ und Dativ bereits vollzogen ist. So weicht im Paradigma der Singularformen lediglich der Genitiv von den übrigen Wortformen ab, indem er das Flexionssuffix *-s* verlangt. Die EN, in althochdeutscher Zeit noch analog zu den Appellativen in die verschiedenen Flexionsklassen eingegliedert, bilden somit heutzutage eine eigene Deklinationsklasse. Auch der Plural, dessen Auftreten aufgrund der Monoreferenz-Eigenschaft eher selten ist, wird bei EN, die nicht bereits auf das Phonem /s/ enden, ausschließlich mit dem Flexiv *-s* gebildet. Indem es an den EN affigiert, ohne ihn zu verändern, schont dieses Suffix den Namenkörper und macht ihn „in jedem Kontext sofort wiedererkennbar“ (**nuebling12**). Dies stellt sicherlich insofern einen Vorteil für automatische EN-Erkennung dar, als die reduzierten Paradigmen zu weniger unbekannten Wortformen führen.
Die besonderen Eigenschaften der EN auf dem Gebiet der Wortbildung erscheinen für die maschinelle EN-Erkennung hingegen nur bedingt hilfreich. So gibt es im Deutschen zwar einige für bestimmte EN-Klassen typische und produktive onymische Suffixe wie beispielsweise *-ien* für Ländernamen, jedoch treten sie nicht derartig regelhaft auf, dass sie Programmen bei der Klassifizierung vollkommen zuverlässige Hinweise böten. Auch sind Sprachverarbeitungsprogramme nicht ohne weiteres in der Lage, Komposita auf semantische Kongruenz zu überprüfen, von der EN häufig abweichen, wie das Beispiel des Ortsnamens *Königswinter* veranschaulicht. Hier könnte ein Programm vermutlich nur aus dem Kontext, beispielsweise durch eine vorangehende Präposition oder durch einen Namenslisteneintrag, ableiten, dass es sich um einen EN handelt und sich den Umstand, dass die dieses Kompositum konstituierenden Appellative in dieser Kombination keine sinnvolle Bedeutung ergeben, nicht zu Nutze machen.

2.2.3 Morphosyntaktische Besonderheiten

Etwas markanter und somit womöglich auch für die Erkennungsprogramme von größerer Relevanz sind die morphosyntaktischen Besonderheiten von EN. Mit der Monoreferenz der EN geht die Eigenschaft der inhärenten Definitheit einher, weshalb EN teilweise keinen Definitartikel benötigen. Dies trifft auf Städte-, die meisten Länder- und auch die Personennamen zu. Straßen-, Gewässer- und Gebirgsnamen tragen hingegen einen festen definiten Artikel (*der Rhein*, *die Goethestraße*, *die Pyrenäen*) und in vielen süddeutschen Varietäten ist auch ein Definitartikel vor Personennamen der Regelfall. Dass Artikellosigkeit und Artikelgebrauch somit, von Ausnahmen abgesehen, ganze EN-Klassen charakterisieren, könnte für die automatische Einteilung von Onym-Instanzen in dieselben durchaus hilfreich sein.

2.2.4 Graphematische Besonderheiten

In den meisten Sprachen, die das lateinische Alphabet benutzen, ist ein großer Anfangsbuchstabe in einer Zeichenkette ein guter Hinweis darauf, das es sich bei ihr um einen EN handelt. Im Deutschen jedoch hat sich die Großschreibung zwischen 1560 und 1710 von den EN auf die gesamte Substantivkategorie ausgebreitet (vgl. **nuebling12**), weshalb sie kein Kriterium für die Unterscheidung von EN und Appellativen darstellt. Dies ist sicherlich ein wichtiger Grund für das Gefälle zwischen der Performanz von EN-Erkennern bei der Annotation von deutschen im Vergleich zu der von englischen Daten beispielsweise. Auch der Umstand, dass EN nicht im eigentlichen Sinne orthographisch normiert sind, kann die automatische Annotation derselben erschweren. Möchte man das Programm beispielsweise mit einem Lexikon von weit verbreiteten Personennamen versehen, werden die vielen verschiedenen Schreibvarianten diese Namensliste um einen beachtlichen Faktor vervielfachen. So ist mit *Schmidt*, *Schmid* und *Schmitt* nur ein Teil der möglichen Schreibweisen dieses häufigen Familiennamens aufgeführt. Auf der anderen Seite kann diese Unnormiertheit aber auch Hinweise für eine richtige Kategorisierung geben. Die besonderen Graphemkombinationen, die EN häufig aufweisen und die mit den bereits in Abschnitt 2.2.1 erwähnten Phonemkombinationen korrespondieren, sowie die häufigere Verwendung von sonst seltenen Graphemen wie <c, q, v, x, y> können unter Umständen die Erkennung eines EN erleichtern.

2.2.5 Zum Nutzen der Eigennamenerkennung

Es ist deutlich geworden, dass Onyme sich zwar in vielerlei Hinsicht von anderen Substantiven unterscheiden, dass es sich bei diesen Unterschieden jedoch größtenteils um prototypische Charakteristiken handelt, welche in der Regel nicht auf alle EN-Klassen zutreffen. Folglich lassen sich keine vollkommen sicheren Indikatoren für das Vorliegen eines Onyms angeben, was eine intensionale Definition unmöglich macht. Aufgrund der Fülle an existierenden EN, ständig hinzukommenden Neubildungen und der Vielzahl an Onymen, die ein appellativisches Homonym haben, ist eine extensionale Aufzählung ebenso wenig möglich. Diese Definitionsschwierigkeiten stellen selbstverständlich auch ein Hindernis für die automatische EN-Erkennung dar. Gleichzeitig ist jedoch gerade für das bessere Verständnis dieser Kategorie die genauere Untersuchung ihrer Verhaltens- und Verwendungsweisen notwendig. Dies ist vor allem unter der Berücksichtigung des Umstands, dass sich die aufgeführten Eigenschaften der EN größtenteils in einem diachronen Prozess herausgebildet haben und sich auch synchron weiterhin im Wandel befinden, relevant. In vielen Bereichen der Onomastik fehlen bisher einschlägige Forschungsergebnisse zu gewissen Themen. Zur Beantwortung der Forschungsfrage beispielsweise, ob und wie weit bei Onymen die Deflexion im Genitiv und Plural vorangeschritten ist, „stehen repräsentative Untersuchungen noch aus“ (**nuebling12**). Es ist offensichtlich,

dass Introspektion und Akzeptabilitätsstudien gerade für ein solches, am Sprachgebrauch interessiertes Forschungsunterfangen keine Alternative zu einer fundierten Korpusstudie darstellen. Dies setzt jedoch eine verlässliche EN-Annotation in der Datengrundlage voraus, weshalb man bei der maschinellen Aufbereitung von Korpora der Performanz der EN-Erkennungsprogramme eine besondere Aufmerksamkeit zukommen lassen sollte.

3 Eigennamenerkennung für das Deutsche

3.1 Ausgangssituation

Die Anzahl frei zugänglicher EN-Erkenner ist für das Deutsche im Gegensatz zum Englischen stark begrenzt. So bin ich bei meiner Recherche lediglich auf drei Klassifizierer gestoßen, die speziell für die EN-Kennzeichnung im Deutschen entwickelt wurden. Dies liegt vermutlich nicht zuletzt daran, dass manuell annotierte Daten, die als Trainingsmaterial genutzt werden könnten, für das Deutsche nur in geringem Ausmaß vorhanden sind. Das einzige mit EN annotierte Korpus, welches den Programmentwicklern somit zur Verfügung stand, umfasst 220.000 Tokens und stammt aus dem CoNLL-2003 Shared Task (**tjong03**), einem Wettbewerb, bei dem EN-Erkenner für das Englische und das Deutsche entwickelt wurden. Die EN darin wurden an der Universität Antwerpen nicht nur gekennzeichnet, sondern zusätzlich in die Unterkategorien Personennamen (*PER*), Ortsnamen (*LOC*), Organisationsnamen (*ORG*) und eine Restklasse (*MISC*) eingeteilt.
Alle drei EN-Erkenner basieren auf dem Ansatz des überwachten maschinellen Lernens. Das heißt, die Algorithmen nutzen zunächst eine von Menschen vorgenommene Klassifizierung als Grundlage für die Generalisierung über bestimmte morphologische, morphosyntaktische und kontextuelle Eigenschaften von EN-Klassen. Sie werten also die aus der Frankfurter Rundschau stammenden, von Hand annotierten CoNLL-2003-Daten aus und wenden daraufhin die daraus abgeleiteten Regeln auf unannotierten Text an. Um die so entstandene Merkmalsrepräsentation der EN und Nicht-EN noch zu ergänzen und die Performanz somit zu steigern, wurden die beiden Programme für das Deutsche ähnlich wie jene, die beim CoNLL-2003 Shared Task entworfen wurden, mit zusätzlichen Lernmechanismen trainiert, die ich im Folgenden kurz beschreiben möchte.

3.2 Der Stanford NER für das Deutsche – Clustering

faruqui10 verfolgten mit ihren Klassifizierern den Ansatz der semantischen Generalisierung auf der Basis von Korpora, in denen die EN nicht annotiert wurden. Das Erkennungs-Programm fasst dabei in diesen Textmengen Wörter, die in einem ähnlichen Kontext vorkommen oder morphologische Gemeinsamkeiten aufweisen, in Gruppen, sogenannte Cluster, zusammen. Unter der Annahme, dass die sich im

selben Cluster befindenden EN auch zur selben EN-Klasse gehören, werden diese dann mit dem entsprechenden Label versehen. So kann das Problem des Mangels an mit EN annotierten Trainingsdaten, der vor allem für das Deutsche akut ist, ohne menschlichen Arbeitsaufwand durch nicht-überwachtes maschinelles Lernen zumindest teilweise behoben werden.
Anstatt von Grund auf einen neuen EN-Erkenner zu erstellen, entwickelten die beiden Forscher zwei Klassifizierer für den Stanford NER (**finkel05**), ein zuvor an der Universität Standford entwickeltes EN-Erkennungsprogramm. Dafür ließen sie den Stanford NER zunächst anhand der Trainingsdaten des CoNLL-2003 Shared Tasks Hinweise für die Kategorie EN analysieren, wobei sowohl wortinterne Indikatoren wie Wortform, Großschreibung und Suffixe als auch kontextuelle Hinweise für die vier EN-Klassen berücksichtigt wurden (für eine tabellarische Auflistung der Merkmale vgl. **finkel05**). In einem zweiten Schritt wendeten sie dann das soeben beschriebene Verfahren der semantischen Generalisierung an. Für einen der Klassifizierer führte der Algorithmus ein Clustering auf den ca. 175 Mio. Tokens umfassenden Daten des Huge German Corpus (HGC) durch, bei denen es sich um Zeitungsmaterial handelt. Das Korpus deWaC (**baroni09**), welches als Clustering-Grundlage für den zweiten Klassifizierer diente, besteht hingegen wie das Korpus, auf welchem ich die Programme ausgewertet habe, aus automatisch aus dem Internet extrahierten Daten. Es ist mit 1,71 Mrd. Tokens bedeutend größer als das HGC und umfasst weitaus mehr Domänen als die CoNLL-2003-Daten, jedoch sind Inhalt und Qualität bedingt durch seine Datengrundlage und sein Ausmaß natürlich weniger kontrolliert. Für den Klassifizierer wurde allerdings nicht im gesamten deWaC geclustert, sondern nur ein Teil mit der Größe des HGC verwendet.

3.3 SemiNER – Clustering und Namenslisten

Im Gegensatz zu Faruqui und Padó programmierten **chrupala10** ihren EN-Erkenner SemiNER von Grund auf. Ihr Algorithmus berücksichtigt jedoch ebenso wie der des Stanford NER wortinterne und kontextuelle Evidenz (für eine tabellarische Auflistung vgl. **chrupala10**) und für den ersten überwachten Lernschritt verwendeten sie dieselben Trainingsdaten des CoNLL-2003 Shared Tasks wie Faruqui und Padó. Auch verfolgten sie einen ähnlichen Clustering-Ansatz, wofür der 34 Mio. Tokens umfassende deutschsprachige Teil des multilingualen Korpus der European Corpus Initiative (ECI)[2] die Grundlage bildete. Jedoch wurden dabei morphologische Indikatoren zunächst nicht berücksichtigt, sondern die Wörter ausschließlich nach ihrer kontextuellen Verteilung klassifiziert. Erst im Nachhinein wurden dann anhand der Gemeinsamkeiten der sich im selben Cluster befindenden Wörter weitere lokale Eigenschaften als Indikatoren für verschiedene EN-Klassen analysiert.
Mit der Auswertung von Wikipedia-Einträgen verwendeten Chrupała und Klakow noch eine zusätzliche Informationsquelle. Häufig enthalten gerade von EN-Denotaten

[2] http://www.elsnet.org/eci.html, letzter Zugriff am 29.06.2013.

handelnde Wikipedia-Einträge eine sogenannte Infobox, die aus einer Tabelle mit wichtigen Eckdaten besteht. Die Artikel zu Russland oder Vietnam enthalten beispielsweise ein Informationskästchen mit Angaben zu den Attributen *Amtssprache*, *Hauptstadt*, *Staatsform*, etc. Die Infoboxen zu Siemens oder Nestlé hingegen listen *Rechtsform*, *Gründung*, *Mitarbeiter* u.ä. auf. Aus allen Wikipedia-Artikeln, die ein solches Informationskästchen enthielten, extrahierten die beiden Forscher die darin aufgeführten Label zusammen mit der Artikelüberschrift. Diese Datengrundlage ermöglichte es dem Programm, Korrelationen zwischen den in der Infobox enthaltenen Attributen und der EN-Klasse, zu der das Artikelthema gehört, herzustellen und daraus wiederum Erkennungsregeln zu generieren. Bei der Klassifizierung sollte der SemiNER somit auf eine größere Anzahl an bereits bekannten EN und eine umfangreichere Eigenschaftenrepräsentation derselben zurückgreifen können.

3.4 Performanz des Stanford NER und des SemiNER

Vermutlich aufgrund der allgemeinen Substantivgroßschreibung und der für das Deutsche charakteristischen morphologischen Komplexität ist bei dem Vergleich der CoNLL-2003-Programme ein starkes Performanzgefälle zwischen den englischen und den deutschen Daten zu beobachten. Das üblicherweise verwendete Maß hierfür ist der F_1-Measure, der das harmonische Mittel aus Precision und Recall der Programme darstellt. Während die Precision das Verhältnis von True Positives und False Positives in dem Ergebnis einer Klassifizierung misst, also im Falle der EN-Erkennung angibt, bei wie viel Prozent der als Onym klassifizierten Tokens es sich auch wirklich um solche handelt, gibt der Recall das Verhältnis von True Positives und False Negatives wieder und stellt somit dar, wie viel Prozent der insgesamt im Korpus enthaltenen EN von dem Programm entdeckt wurden. Die Precision eines EN-Erkenners ist damit desto höher, je weniger Subtantive er fehlerhaft als EN klassifiziert. Der Recall hingegen ist umso höher, je kleiner die Anzahl der EN ausfällt, die das Programm nicht erkannt hat. Der F_1-Measure ($F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$) fasst diese beiden Größen zusammen und liegt beim besten Programm des CoNLL-2003 Shared Tasks für das Englische bei 88,8%, für das Deutsche jedoch nur bei 72,4%. Diese Ergebnisse wurden wohlgemerkt auf den CoNLL-2003-Testdaten erzielt, die insofern eine große Homogenität mit den Trainingsdaten aufweisen, als sie aus derselben Zeitung stammen. Auch die Entwickler des SemiNER und jene der beiden Klassifizierer für den Stanford NER haben ihre Programme auf diesen Daten evaluiert.

Tabelle 1: Performanz der Programme auf den Test-Daten des CoNLL-2003 Shared Tasks

Programm	Precision	Recall	F_1-Measure
HGC	86,6	71,2	78,2
deWaC	86,4	68,5	76,4
SemiNER	80,3	69,8	74,7

Wie der Tabelle 1 entnommen werden kann, erzielte der EN-Erkenner von Chrupała und Klakow hierbei einen F_1-Measure von 74,7%. Die Forscher stellten zudem fest, dass das Clustering zu einer wesentlich größeren Performanzsteigerung führte als die Wikipedia-Informationen. Letztere schienen für eine größer werdende Menge an Trainingsmaterial sogar überhaupt keine Verbesserung des F_1-Measures mehr zu bewirken: „For larger sizes of the training set, using infobox features on top of distributional cluster features does not give any further improvements“ (**chrupala10**).
Der deWaC-Klassifizierer von Faruqui und Padó erzielte auf den CoNLL-2003-Daten einen F_1-Measure von 76,4%, während der HGC-Klassifizierer sogar 78,2% erreichte. Sowohl für diese beiden Klassifizierer als auch für den SemiNER gilt, dass der Recall, der jeweils bei ca. 70% liegt, um mindestens 10% schlechter ausfiel als die Precision. Dies lässt sich das Deutsche betreffend auch für die CoNLL-2003-Ergebnisse feststellen, während für das Englische die beiden Maße in den meisten Fällen nicht mehr als 2% voneinander abweichen.
Um ihre Klassifizierer auch auf den CoNLL-2003-Daten weniger ähnlichen Texten zu testen, werteten Faruqui und Padó deren Performanz zusätzlich auf dem Korpus EUROPARL (**koehn05**) aus, welches statt aus Zeitungstexten aus Tagungsberichten des Europäischen Parlaments besteht. Die Klassifizierer büßten dabei jeweils mehr als 10% an F_1-Measure ein und sanken auf 65,3% (deWaC), bzw. 65,6% (HGC). Angesichts dieses Performanzverlusts ist es nun besonders interessant, die Programme auf einem Web-Korpus, welches sowohl schrift- als auch eher spontansprachliches Material und eine große Bandbreite an Domänen enthält, zu evaluieren und zu überprüfen, ob die Ergebnisse überhaupt jene der inhärenten EN-Erkennung des TreeTaggers (**schmid94**) übertreffen.

4 Evaluation der Pogramme

4.1 Datengrundlage: CatTle.de.12

Die Daten für die von mir durchgeführte Evaluation der Programme stammen aus dem CatTle.de.12, welches seinerseits aus dem DECOW2012 (**schaefer12**), einem 9,1 Mrd. Tokens umfassenden gecrawlten Web-Korpus, abgeleitet wurde und sich noch in der Aufbereitungsphase befindet. Der Vorteil davon, das Korpus CatTle.de.12 als Grundgesamtheit für das Testmaterial zu wählen, besteht vor allen Dingen darin, dass es bereits eine Klassifizierung bezüglich einiger Kategorien erfahren hat, die für die Performanz der EN-Erkennung relevant sein könnten. So wurde beispielsweise für jedes Dokument in der Kategorie *Mode* ermittelt, ob es schriftsprachliches (*Written*) oder eher spontansprachliches (*Quasi-Spontaneous*) Material enthält[3]. Da zu erwarten ist, dass in spontansprachlichen Dokumenten orthografische Normen weniger berücksichtigt werden und vor allem die Großschreibung, die ein wichtiges Erkennungsmerkmal für Substantive und damit für potentielle EN darstellt, häufig

[3] Vgl. Kategorisierungsrichtlinien: http://hpsg.fu-berlin.de/cow/files/cowcat2013.pdf, letzter Zugriff am 13.07.2013.

vollkommen weggelassen wird, ist hier mit einer schlechteren Performanz als auf den schriftsprachlichen Daten zu rechnen. Zusätzlich hat die Kategorie *Mode* noch die Ausprägung *Spoken*, zu der hauptsächlich Interviews gehören. Meist sind diese stark editiert, weshalb ich Dokumente dieser Unterkategorie in der späteren statistischen Auswertung zu *Written* zähle. Dokumente, welche mit einem schriftsprachlichen Beitrag beginnen auf den eher spontansprachliche Kommentare folgen, werden im CatTle.de.12 auch noch einmal von den anderen drei Ausprägungen unterschieden und unter *Blogmix* gefasst. Da die Kommentarbeiträge bei den in meiner Stichprobe enthaltenen Dokumenten in der Tokenanzahl meist überwiegen, zähle ich sie für die Auswertung zur Kategorie *Quasi-Spontaneous*.
Ebenfalls einen Einfluss auf die Güte der EN-Erkennung könnte es haben, welcher Wert einem Dokument in der Kategorie *Audience* zugewiesen wurde. Hier wird kodiert, ob das Textdokument für ein allgemeines Publikum verständlich ist (*General*), sich an informierte Leser (*Informed)* oder gar an solche, die in dem Themenbereich des Dokuments eine Berufsausbildung absolviert haben (*Professional*), richtet. Es scheint plausibel, dass in Texten, die an Personen mit Vorkenntnissen adressiert sind, weniger frequente EN auftauchen und die EN-Erkennung an Performanz verliert.
Um zu untersuchen, ob es eine Korrelation zwischen der Performanz der EN-Erkenner und den Werten in den Kategorien *Mode* und *Audience* der jeweiligen Dokumente gibt, habe ich die Stichprobe nach der in Tabelle 2 wiedergegebenen Verteilung im 803 Dokumente umfassenden CatTle.de.12 proportional stratifiziert.

Tabelle 2: Verteilung innerhalb der Kategorien *Mode* und *Audience* im CatTle.de.12 in Prozent. Die Werte in Klammern geben die Verteilung in der Stichprobe an, sofern diese abweichend ist. Steht ein horizontaler Strich, betrug der Anteil im CatTle.de.12 weniger als ein Prozent.

Audience / **Mode**	General	Informed	Professional	**Total**
Written	63 (64)	4	2	**69**
Quasi-Spontaneous	14 (17)	11	2	**27**
Blogmix	3 (0)	–	–	**3**
Spoken	1 (0)	–	–	**1**
Total	**81**	**15**	**4**	**100**

Da die Stichprobe genau 100 Dokumente umfasst, sind die Prozentangaben aus Tabelle 2 mit den absoluten Dokumentzahlen in der Stichprobe zwar identisch, allerdings habe ich, wie oben erwähnt, die Kategorien *Written* und *Spoken* sowie *Quasi-Spontaneous* und *Blogmix* zusammengefasst. Aufgrund dessen wird für die statistische Auswertung von 64 Dokumenten, die als *Written* und *General* eingestuft wurden, und 17 Dokumenten aus der Kategorie *Quasi-Spontaneous* in der Kombination mit *General* ausgegangen. *Blogmix* und *Spoken* sind nach dieser Annahme gar nicht mehr in der Stichprobe vertreten. Diese Veränderungen sind in der Tabelle in Klammern gekennzeichnet.
Mit insgesamt ca. 104.600 Tokens ist die Stichprobe fast doppelt so groß wie das 55.000 Tokens umfassende Test-Korpus des CoNLL-2003 Shared Tasks, wodurch si-

chergestellt werden sollte, dass eine für die statistische Auswertung genügend große Anzahl von EN-Instanzen darin enthalten ist.

4.2 Richtlinien

Um die EN-Erkenner auf ihre Güte zu überprüfen, ist eine manuelle Annotation, die den Idealzustand darstellt und an der die Ergebnisse der Programme gemessen werden können, nötig. Das Kapitel 2 hat jedoch gezeigt, dass eine Definition von EN weder intensional noch extensional möglich ist und auch unter Linguisten nicht immer Einigkeit darüber herrscht, ob es sich bei einem Wort um ein Onym handelt, bzw. darüber, welcher EN-Klasse gewisse Onyme angehören. Dies führt zwangsläufig zu unscharfen Rändern der Kategorien und die zahlreichen Zweifelsfälle, auf die man bei der manuellen Annotation stößt, spiegeln sich in der Vielzahl unterschiedlicher Richtlinien zur EN-Kennzeichnung wider.
Die eigens für die EN-Erkennung konzipierten Programme, die ich im Rahmen dieser Arbeit getestet habe, basieren auf den Trainingsdaten des CoNLL-2003 Shared Tasks, denen Richtlinien zugrunde liegen, die lediglich aus einer Aufzählung von semantischen Unterkategorien für jede der vier EN-Klassen *PER*, *LOC*, *ORG* und *MISC* ohne jeglichen Kommentar bestehen[4]. Einige der semantischen Subkategorien finden sich in mehr als einer EN-Klasse wieder. Hierzu zählen vor allem solche EN, die sowohl als Orts- als auch als Organisationsbezeichnung fungieren können. Während in (1) das *Deutsche Theater* auf das Gebäude referiert, also einen Ort bezeichnet, ist in (2) von der Institution die Rede, weshalb hier eher das Label *ORG* angebracht erscheint.

(1) *Die Fassade des* $[$*Deutschen Theaters*$]_{LOC}$ *muss saniert werden.*

(2) *Das* $[$*Deutsche Theater*$]_{ORG}$ *zeigt ausschließlich interessante Stücke.*

Vermutlich geht die doppelte Nennung einiger Kategorien darauf zurück, dass diese Unterscheidung bei der Annotation berücksichtigt werden soll, jedoch wird dies in den Richtlinien nicht dargelegt.
Auf die Probleme bestehender Richtlinien werde ich im Abschnitt 6.2 noch einmal zurückkommen. Hier soll zunächst deutlich werden, dass die Entscheidung, ob ein Wort als EN zu kategorisieren ist und zu welcher Subklasse es gehört, keinesfalls trivial ist und nicht immer gleich gehandhabt wird. So unterscheiden sich die STTS-Richtlinien[5], die der TreeTagger verwendet, von denen des CoNLL-2003 Shared Tasks nicht allein dadurch, dass keine weitere Unterklassifizierung der EN vorgenommen wird. Vielmehr liegt ihnen eine vollkommen andere Auffassung darüber zugrunde, was in einem Korpus als EN zu taggen ist. Während die CoNLL-2003-Richtlinien ein eher großzügiges Namenskonzept vorschlagen, werden nach den STTS-Richtlinien fast ausschließlich idiosynkratische EN mit dem Label *NE* für *Named Entitiy* ver-

[4] http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt, letzter Zugriff am 13.07.2013.
[5] http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, letzter Zugriff am 13.07.2013.

sehen. In komplexen EN sollen hingegen die genuinen Wortarten zugewiesen werden. So wären beispielsweise nach den Richtlinien des CoNLL-2003 Shared Task in *Deutsches Theater* beide Tokens als *ORG*, bzw. je nach Verwendung als *LOC* zu klassifizieren, während nach STTS hier *Deutsches* als Adjektiv und *Theater* als Substantiv gekennzeichnet und das Label *NE* überhaupt nicht verwendet würde. Auch Determinativkomposita, wie sie häufig bei Straßennamen vorliegen, und Produktnamen werden nach STTS nicht zu den EN gezählt. Diese grundlegenden Unterschiede in den Richtlinien machen eine Auswertung anhand des selben Maßstabs selbstverständlich unmöglich. Daher erstellte ich für die Evaluation der Programme zwei Goldstandards, wobei einer den STTS-Richtlinien folgt und der andere der Klassifikation des CoNLL-2003 Shared Tasks. Hierbei impliziert das umfassendere Namenskonzept des CoNLL-2003 Shared Tasks von vornherein, dass der TreeTagger insgesamt eine kleinere Menge von Tokens als EN klassifiziert, als die anderen Programme. Unabhängig von der Gesamtmenge der EN soll jedoch zunächst die Performanz der Programme in den dafür üblichen Maßen Precision, Recall und dem harmonischen Mittel daraus verglichen werden.

4.3 Evaluationsergebnisse

In Abschnitt 3.4 wurde bereits bemerkt, dass die EN-Erkennung für das Deutsche selbst auf den Trainingsdaten sehr ähnlichem Testmaterial wesentlich weniger performant ist, als es für das Englische die Regel ist. Der von Faruqui und Padó durchgeführte Test ihrer Programme auf den EUROPARL-Daten lieferte zusätzlich um etwa 10% schlechtere Ergebnisse. Somit ist es nicht verwunderlich, dass die Performanz auf den CatTle.de.12-Daten, welche sowohl intern als auch zu den Trainingsmaterialien des CoNLL-2003 Shared Tasks eine starke Inhomogenität aufweisen, weiter abfällt. Wichtig ist zu erwähnen, dass in der Auswertung solche Tokens, die zwar korrekterweise als EN erkannt, jedoch der falschen Subklasse zugeordnet wurden, genauso als Fehler angesehen werden, wie gar nicht erkannte EN. Dies führt dazu, dass in der Fehlerauswertung falsch kategorisierte EN doppelt gezählt werden, da sie sowohl für die fälschlich vergebene Kategorie als False Positive als auch für ihre eigentliche Kategorie als False Negative gelten. An späterer Stelle wird zum Vergleich dargestellt, wie sich die Messwerte für die Performanz verändern, wenn man die richtige Unterkategorisierung der EN nicht mehr berücksichtigt.
Die Tabelle 3 zeigt, dass sowohl der SemiNER, als auch der deWaC- und der HGC-Klassifizierer einen um mindestens 20% niedrigeren F_1-Measure auf den Internetdaten erzielen, als auf den Zeitungsmaterialien des CoNLL-2003 Shared Tasks.
Der SemiNER büßt sogar 27,5% ein und liefert gleichzeitig mit 47,2% F_1-Measure das schlechteste Ergebnis, während der TreeTagger von allen vier Programmen mit 67,8% den höchsten F_1-Measure erzielt. Allerdings ist daran zu erinnern, dass die STTS-Richtlinien eine sehr sparsame EN-Annotation vorsehen. So gehen die Berechnungen für die drei eigens für die EN-Erkennung konzipierten Programme auf Basis der

Tabelle 3: Performanz der Programme auf den Daten des CatTle.de.12 im Vergleich zu jenen des CoNLL-2003 Shared Tasks

TreeTagger			
EN-Klasse	Precision	Recall	F_1-Measure
Total	**69,5**	**66,1**	**67,8**
deWaC			
EN-Klasse	Precision	Recall	F_1-Measure
PER	83,6	65 4	73,1
LOC	73,2	54,2	62,3
ORG	45,5	38,8	41,9
MISC	63,6	19,0	29,3
Total	**69,8**	**45,5**	**55,2**
Total (CoNLL)	**86,4**	**68,5**	**76,4**
Performanzverlust	**16,6**	**23,0**	**21,2**
HGC			
EN-Klasse	Precision	Recall	F_1-Measure
PER	85,0	60,4	70,6
LOC	76,5	47,5	58,6
ORG	47,4	37,1	41,6
MISC	64,2	16,7	26,5
Total	**71,6**	**41,6**	**52,6**
Total (CoNLL)	**86,6**	**71,2**	**78,2**
Performanzverlust	**15,0**	**29,6**	**25,6**
SemiNER			
EN-Klasse	Precision	Recall	F_1-Measure
PER	70,5	60,5	65,2
LOC	54,6	45,6	49,7
ORG	33,5	36,6	34,8
MISC	52,2	17,6	26,3
Total	**55,1**	**41,3**	**47,2**
Total (CoNLL)	**80,3**	**69,8**	**74,7**
Performanzverlust	**25,2**	**28,5**	**27,5**

CoNLL-2003 Shared Task-Richtlinien nach meiner Annotation von insgesamt 5758 EN in der Stichprobe aus, während im Goldstandard für den TreeTagger lediglich 3585 Tokens das Label *NE* erhielten.

Interessanterweise erreicht der deWaC-Klassifizierer, der zuvor einen um 1,8% niedrigeren F_1-Measure aufzeigte als der HGC-Klassifizierer, auf den CatTle.de.12-Daten nun um 2,8% bessere Ergebnisse als letzterer. Zwar ist die Precision des HGC-Klassifizierers noch immer höher, der deWaC-Klassifizierer erreicht jedoch einen um 3,9% höheren Recall-Wert. Möglicherweise ist dies auf den Umstand zurückzuführen, dass es sich sowohl beim deWaC als auch beim CatTle.de.12 um Web-Korpora handelt. Somit weisen die Clustering-Daten des deWaC-Klassifizierers eine größere Ähnlichkeit zu jenen des CatTle.de.12 auf, was den Umgang mit einer solchen Bandbreite an Registern und Domänen, die automatisch erstellte Web-Korpora stets enthalten, etwas erleichtern könnte. Allerdings ist dies nur eine Mutmaßung, die im Rahmen dieser Arbeit nicht überprüft wurde und weiterer Forschung bedürfte.

Für alle vier Programme lässt sich feststellen, dass der Recall auf den CatTle.de.12-Daten wesentlich stärker absinkt als die Precision. Zwar liegt er im Allgemeinen bei der EN-Erkennung im Deutschen meist unter dem Wert der Precision, jedoch nimmt der Abstand hier noch einmal zu und erreicht beim HGC-Klassifizierer sogar 30%. Dass somit keines der drei Programme mehr als 45,5% aller in der Stichprobe vorhandenen EN erkennt, schränkt deren Nutzbarkeit für Forschungszwecke selbstverständlich stark ein. Zwar ist auch die Precision nicht zufriedenstellend, allerdings können erhobene Daten, wenn es auch einigen Aufwand kostet, in einem zweiten Arbeitsschritt stets von den False Positives bereinigt werden. Die False Negatives jedoch bleiben unauffindbar, weshalb die Ergebnisse besonders beklagenswert sind.
Bei jenen Programmen, die eine weitere Einteilung der EN in Unterklassen vornehmen, lässt sich zudem ein starkes Performanzgefälle zwischen den einzelnen Kategorien beobachten. So sinkt der F_1-Measure von *PER* über *LOC* und *ORG* jeweils um nicht weniger als 10% und erreicht bei der Kategorie *MISC* einen Tiefpunkt, indem er für kein Programm 30% überschreitet. Dies ist sicherlich auf die spezifischen Eigenschaften der EN in jeder der vier Subklassen zurückzuführen. Während bei der Annotation von Personennamen noch wenig Zweifelsfälle auftauchen, weist die Kategorie *LOC* bereits weniger scharfe Ränder auf. Zum einen können einige EN, wie in Abschnitt 4.2 bereits dargelegt wurde, sowohl als Orts- als auch als Institutionsname verwendet werden. Zum anderen umfasst die Kategorie *LOC* sehr viele Gattungs-EN und komplexe Namen, die häufig Appellative enthalten und somit außerhalb des Kontextes ambig sein können. So haben beispielsweise die komplexen Toponyme *Dorf Mecklenburg* und *Burg Olbrück* dadurch zu False Negatives geführt, dass *Dorf* und *Burg* als Appellativ klassifiziert wurden, wenngleich sie hier Teil des EN sind. Darüber hinaus enthielt die Stichprobe auch mehrteilige oder kompositionelle Ortsbezeichnungen, die ausschließlich aus Appellativen bestehen. Beispiele hierfür sind *Alte Feuerwache* und *Fliegerdenkmal*. Hier ist es auch für menschliche Annotatoren selbst mit Kontextkenntnissen zunächst nicht ganz einfach zu entscheiden, ob es sich um eine appellativische Bezeichnung handelt oder ob hier tatsächlich ein EN vorliegt. Die Großschreibung des Adjektivs in *Alte Feuerwache* ist zwar ein Hinweis auf dessen Onymstatus, beim Fliegerdenkmal jedoch hilft dieses Kriterium aufgrund der allgemeinen Substantivgroßschreibung im Deutschen nicht weiter. Auch Organisationsnamen können vollkommene appellativische Transparenz aufweisen. Betrachtet man zum Beispiel die Belege *Landesamt für Kultur- und Denkmalpflege* oder *Büro für Architektur und Planung* aus dem CatTle.de.12, wird schnell klar, dass es für ein Programm äußerst schwierig ist, zu erkennen, dass es sich bei dieser Sequenz aus Appellativen um einen EN handelt.
Da die Kategorie *MISC* eine Art Restklasse für EN, die weder Personen-, noch Orts- oder Organisationsnamen sind, darstellt, handelt es sich bei ihrem Inhalt um eine sehr inhomogene Menge von Wörtern. Ähnliche Probleme wie in der Kategorie *LOC* und *ORG* dürften beispielsweise die unter der Kategorie *MISC* gefassten Film- und Buchtitel verursachen, die häufig ähnlich syntagmatisch sind, wie die oben genann-

ten Beispiele für Institutionsnamen. Ebenfalls scheint die Erkennung von adjektivischen oder nominalen EN-Derivationen wie *dänisch* oder *Singapur-Besucher*, die nach den CoNLL-2003-Richtlinien ebenfalls unter *MISC* gruppiert werden, Schwierigkeiten zu bereiten.
Die aufgeführten Probleme stellen jedoch nur einige der vielen Faktoren dar, die die große Anzahl falscher Klassifikationen erklärbar machen. Um herauszufinden, welche spezifischen kontextuellen und wortinternen Eigenschaften zu Falschklassifikationen führen, und unter Berücksichtigung dieser Einflussfaktoren dann Maßnahmen für die Performanzsteigerung ableiten zu können, ist zunächst eine genauere Untersuchung der fehlerhaften Annotationen nötig. Zu diesem Zweck habe ich für die vier Programme eine statistische Fehleranalyse durchgeführt, deren Ergebnisse nach einer kurzen Beschreibung meines Vorgehens im Folgenden dargestellt werden sollen.

5 Fehlerklassifikation

5.1 Vorgehen: Regressionsanalyse

Da mit Ausnahme des TreeTaggers keines der Programme weniger als 4000 Falschklassifikationen vorgenommen hatte, konnte ich für die Analyse nicht alle Fehler im Detail betrachten und habe stattdessen für jedes der vier Programme Zufallsstichproben von je 100 False Positives und False Negatives betrachtet. Um Hypothesen darüber aufzustellen, welche Eigenschaften eines Wortes eine Falschklassifikation desselben durch die Programme wahrscheinlicher machen, habe ich die Belege zunächst einzeln durchgesehen und sowohl ihre graphematischen und morphosyntaktischen Eigenschaften als auch ihren Kontext genauer betrachtet. Um diese zunächst explorativ gewonnenen Hypothesen daraufhin empirisch zu überprüfen, bot sich das Verfahren der logistischen Regression an, welches den Einfluss mehrerer unabhängiger Variablen auf eine abhängige Variable zu schätzen hilft. Die Unabhängigkeit der Einflussgrößen muss dabei dadurch gegeben sein, dass keine Ausprägung der einen Variable eine bestimmte Ausprägung einer anderen impliziert. Im einfachsten Fall werden unter der abhängigen Variablen zwei sich gegenseitig ausschließende, sogenannte komplementäre Ereignisse betrachtet und es kann dann untersucht werden, ob und in wiefern die unabhängigen Größen das Eintreten des einen oder des anderen Komplementärereignisses beeinflussen. Bei dieser Fehleranalyse wäre demnach das eine Komplementärereignis ein True Positive, also ein richtig erkannter EN, und das andere ein False Negative, ein EN, der nicht als solcher erkannt wurde. Da es allerdings nicht nur interessant schien, zu untersuchen, aus welchen Gründen die Programme einen EN nicht erkennen, sondern ich auch herausfinden wollte, was die Programme dazu führt, ein Wort fälschlicherweise als EN zu klassifizieren, habe ich zusätzlich in einem weiteren Regressionsverfahren die False Positives den True Negatives gegenübergestellt. Zu diesem Zweck habe ich für jedes Programm die Zu-

fallsstichprobe aus den False Positives mit einer ebenfalls 100 Belege umfassenden Zufallsstichprobe aus den True Negatives zusammengeführt und bin analog mit den False Negatives und True Positives vorgegangen.
Die unabhängigen Variablen, bei denen es sich um die verschiedenen vermeintlichen Einflussgrößen handelt, bestanden aus den Features, die die Algorithmen bei der Klassifikation eines Wortes berücksichtigen, ergänzt durch die weiteren möglichen Einflussfaktoren, die mir bei der ersten Durchsicht der Fehler aufgefallen waren. Dabei ist ein Regressionsmodell mit 14 unabhängigen Variablen für die EN und 13 für die Nicht-Onyme entstanden. Die unterschiedliche Anzahl kommt dadurch zustande, dass ich für die EN auch kodiert habe, ob sie ein nicht-onymisches Homonym besitzen, was sich für die Nicht-EN erübrigt. Vor der statistischen Auswertung habe ich zunächst den je 400 Belegen pro Programm einen Wert für alle 14 Einflussgrößen zugewiesen. Unter anderem wurde kodiert, ob sich im Links- oder Rechtskontext des zu Klassifizierenden Tokens EN befinden, aber auch die wortinternen Eigenschaften wie Kleinschreibung oder das Vorhandensein von Sonderzeichen oder Flexions- und Derivationssuffixen wurden gekennzeichnet. Zusätzlich habe ich für jeden Beleg markiert, welchen Wert das Dokument, aus dem er stammt, in der Kategorie *Mode* und *Audience* bei der dieser Arbeit vorausgegangenen Klassifizierung der CatTle.de.12-Daten erhalten hatte. Da sich von diesen unabhängigen Variablen nie alle, aber stets einige als signifikant erwiesen, habe ich die Modelle dann für jedes Programm auf die tatsächlich Einfluss nehmenden unabhängigen Variablen reduziert und sie anschließend mithilfe statistischer Tests auf ihre Güte überprüft. Eine detaillierte Darstellung der für das Regressionsverfahren relevanten Werte und der Analyse der Modellgüte finden sich in Tabelle 7 im Anhang. Im nun folgenden Abschnitt sollen die Ergebnisse lediglich anhand der sogenannten Odds-Ratios, die in der Tabelle im Anhang zusätzlich mit ihrem 95%-Konfidenzintervall aufgeführt sind, eingehender besprochen werden.

5.2 Regressionsergebnisse

5.2.1 Die verschiedenen Einflussgrößen

Die Tabelle 4 zeigt eine Zusammenstellung der signifikanten Ausprägungen der Einflussgrößen und gibt mit den Odds-Ratios wieder, um welchen Faktor der Wert einer unabhängigen Variable die Fehlerwahrscheinlichkeit steigert. Bei den Odds-Ratios handelt es sich um den Quotienten aus der Wahrscheinlichkeit für eine Falschklassifikation im Verhältnis zu der Wahrscheinlichkeit einer richtigen Kategorisierung. Liegt der Wert für eine Einflussgröße oberhalb der Eins, erhöht diese die Fehlerwahrscheinlichkeit, während sie sie verringert, wenn der Wert zwischen null und eins liegt. So bedeutet beispielsweise der ganz links oben in der Tabelle stehende Wert 4,4, dass ein dem zu klassifizierenden Token direkt vorangehender EN dessen fälschliche Annotation als EN um etwas mehr als das Vierfache erhöht. In der linken Tabellenhälfte sind die Odds-Ratios für die False Positives angegeben, also

fälschlicherweise als EN gekennzeichnete Tokens, und in der rechten jene für die nicht-erkannten EN. Enthält eine Zelle einen horizontalen Strich, so hat die in der dazugehörigen Zeile aufgeführte Ausprägung der Einflussgröße für das in der dazugehörigen Spalte stehende Programm keinen signifikanten Einfluss auf die Fehlklassifikation. Als Signifikanzniveau wurde wie in statistischen Testverfahren üblich α = 0,05 gewählt.

Tabelle 4: Odds-Ratios der Einflussgrößen in ihren Ausprägungen für die vier Programme im Vergleich

	False Positives				**False Negatives**			
	TreeTagger	**deWaC**	**HGC**	**SemiNER**	**TreeTagger**	**deWaC**	**HGC**	**SemiNER**
Einflussgröße	**False Positives**				**False Negatives**			
Kontexteigenschaften								
EN an Position -1	4,4	127,0	23,0	199,4	0,5	0,2	0,4	0,1
EN an Position +1	–	23,1	24,8	28,6	0,3	0,3	0,6	0,03
Def. Artikel an Position -1	–	4,4	–	–	4,7	2,6	–	2,9
$\emptyset$ an Position -1	–	–	–	5,9	–	–	–	2,4
Lokale Eigenschaften								
Abkürzung	20,4	–	–	146,6	–	–	–	–
Sonderzeichen	24,1	12,4	–	–	6,5	4,7	3,8	–
ausschließlich Minuskeln	0,1	0,001	0,02	–	7,9	9,8	5,5	12,7
ausschließlich Majuskeln	–	–	20,4	–	–	–	–	–
Flexionssuffix	0,1	0,3	–	–	–	–	–	–
Derivationssuffix	0,3	–	–	–	6,3	–	–	–
Ambig					–	4,8	13,7	6,8
Fremdsprachlich	–	–	33,7	48,7	–	8,4	–	–
Registerinformationen								
Mode: Quasi-Spontaneous	2,0	2,6	–	4,6	0,5	2,0	1,8	–
Audience: Informed	–	0,3	–	–	–	–	–	–

Wie bereits erwähnt, wurden sowohl kontextuelle als auch lokale Eigenschaften der Tokens in die Analyse miteinbezogen. Zusätzlich bot es sich an diesem Punkt der Evaluation an, zu überprüfen, ob das sprachliche Register einen Einfluss auf die Performanz der EN-Erkenner ausübt, weshalb auch die Metadaten aus der Kodierung des CatTle.de.12 unter *Registerinformationen* als unabhängige Variablen angeführt sind. Die den Kontext betreffenden Einflussgrößen kodieren abgesehen von adjazenten EN sowohl das Vorhandensein von definiten Artikeln als auch von Interpunktions- und Satzendezeichen, in der Tabelle gekennzeichnet durch das Symbol der leeren Menge, in unmittelbarer Präzedenz des zu klassifizierenden Tokens. Die unabhängigen Variablen zu den lokalen Eigenschaften beschreiben einerseits morphosyntaktische Besonderheiten wie das Vorhandensein von Flexions- und Derivationssuffixen oder ob es sich bei einem Token um eine Abkürzung handelt. Andererseits kennzeichnen sie auch graphematische Auffälligkeiten wie die ausschließliche Verwendung von Majuskeln, bzw. Minuskeln oder das Vorhandensein von Son-

derzeichen im Schriftbild. Außerdem wurde für jedes Token festgehalten, ob es sich bei ihm um fremdsprachliches Material handelt, da dies auch beim POS-Tagging eine häufige Fehlerquelle darstellt (**giesbrecht09**). Für die aus Onymen bestehenden Stichproben wurde mit der Variable *Ambig*, die für jeden EN kennzeichnet, ob zu ihm ein nicht-onymisches Homonym existiert oder nicht, zusätzlich ein semantisches Kriterium berücksichtigt.
Für die verschiedenen Programme haben sich interessanterweise sowohl in Bezug auf die Art der Einflussgrößen als auch auf deren Anzahl sehr unterschiedliche Modelle als passend erwiesen. Während sich für den HGC-Klassifizierer für die False Positives lediglich fünf und für die False Negatives sechs Ausprägungen der Variablen als signifikant erwiesen haben, sind es bei dem des deWaC jeweils acht. Dennoch lassen sich in der Tabelle einige Einflussgrößen erkennen, die für alle oder die meisten der Programme eine Rolle spielen. Im Folgenden sollen die signifikanten Ausprägungen der unabhängigen Variablen eingehender besprochen und, sofern möglich, mit den bei der explorativen Erstdurchsicht der Fehler gemachten Beobachtungen in Bezug gesetzt werden.

5.2.2 Einfluss der kontextuellen Variablen

Es ist sofort zu erkennen, dass das Vorhandensein eines EN an linker oder rechter Position des zu klassifizierenden Tokens bei fast allen Programmen die Wahrscheinlichkeit für einen False Positive um einen beträchtlichen Faktor erhöht. Hier kann es sich bei dem adjazenten, von den Programmen als EN klassifizierten Token sowohl um einen True Positive als auch um einen False Positive handeln. Nicht selten iterieren sich somit Fehlklassifikationen, zum Beispiel in solchen Fällen, in denen ein Personen- oder ein Ortsname in einem Organisationsnamen enthalten ist und dazu führt, dass der gesamte EN-Komplex fälschlicherweise mit *PER*, bzw. *LOC* gekennzeichnet wird. Gleichzeitig senken EN in der Umgebung eines Onyms die Wahrscheinlichkeit, dass dieses nicht als solches erkannt wird, was dadurch deutlich wird, dass die Odds-Ratios für die False Negatives alle unter eins liegen. Dies spiegelt sich unter anderem in der bei der ersten Datendurchsicht gemachten Beobachtung wider, dass die Programme Personennamen besonders gut zu erkennen scheinen, wenn Vor- und Nachname in direkter Adjazenz im Korpus auftreten. In Anbetracht der Werte der Odds-Ratios und der Signifikanz für fast alle Programme sowohl im Regressionsmodell für die False Positives als auch für die False Negatives kann hier sicherlich von einem sehr einflussreichen Faktor für das Entstehen von Fehlern gesprochen werden.
Jedoch sind EN nicht die einzigen Instanzen, die im Linkskontext einen Einfluss auf die Ausprägung der abhängigen Variable nehmen. Vielmehr scheint auch ein definiter Artikel die Fehlerwahrscheinlichkeit sowohl für die False Positives als auch für die False Negatives zu erhöhen. Die Interpretation dieses Ergebnisses ist nicht ganz einfach. Da die automatisch lernenden EN-Erkenner anhand großer Textmengen statistisch auswerten, was typischerweise im Linkskontext eines EN vorkommt,

würde man erwarten, dass eine kontextuelle Variable stets auf der einen Seite die Fehlerwahrscheinlichkeit senkt und sie auf der anderen Seite erhöht, nicht jedoch beiderseits die Fehlerwahrscheinlichkeit steigert. Bei dem Auftreten von Onymen im Linkskontext ist dies zum Beispiel der Fall und es lässt sich schlussfolgern, dass einem EN häufig andere als EN klassifizierte Tokens direkt vorangehen. Für die definiten Artikel lässt sich jedoch lediglich vermuten, dass die beidseitig über eins liegenden Werte darauf zurückzuführen sind, dass einige EN-Klassen, wie in Abschnitt 2.2.3 dargelegt wurde, keinen Definitartikel nehmen, während explizite Definitheit bei anderen die Regel ist. Da beispielsweise Personennamen in den meisten Varietäten des Deutschen keinen definiten Artikel nehmen, könnte die Verwendung desselben in Verbindung mit einem Anthroponym zu dessen Falschklassifikation beitragen, sodass ein False Negative entstünde. Andersherum könnte ein False Positive in solchen Fällen wahrscheinlicher werden, in denen es sich bei dem zu klassifizierenden Token um ein ambiges Substantiv handelt. So könnte *Marktplatz* oder *Landschulheim* in Verbindung mit einem definiten Artikel eher als EN gewertet werden als bei Verwendung eines indefiniten Artikels. Dass solche Appellative auch häufig in definiten Nominalphrasen einem tatsächlichen EN als Apposition vorangehen, wie es in *das Hotel [Linzer Hof]*$_{LOC}$ der Fall ist, könnte ebenfalls dazu führen, dass statt der angegebenen richtigen Kennzeichnung auch *Hotel* das Label *LOC* erhält. Dies ist jedoch lediglich eine spekulative Hypothese, für deren Rechtfertigung eine hier aus zeitlichen Gründen nicht durchführbare qualitative Auswertung einer umfangreichen Menge an Fehlerbelegen, die einen definiten Artikel im Linkskontext aufweisen, nötig wäre.

Für den SemiNER hat sich zusätzlich ein leerer Linkskontext, in der Tabelle durch das Symbol ∅ gekennzeichnet, als signifikant Einfluss auf die Fehlerentstehung nehmend erwiesen. Die Werte zeigen an, dass sowohl False Positives als auch False Negatives am Satzanfang oder auf Interpunktion folgend mit einer größeren Wahrscheinlichkeit auftreten. Dies ist insofern kein überraschendes Ergebnis, als ein Satzanfang und Interpunktion kaum Informationen darüber liefern, ob es sich bei dem darauffolgenden Token eher um einen EN handelt oder nicht. Kontextuelle Hinweise auf die Zugehörigkeit zu einer bestimmten EN-Klasse, wie sie zum Beispiel die Präpositionen für die Kategorie *LOC* darstellen, fehlen ebenfalls und erschweren eine korrekte Erkennung.

In Anbetracht der aufgeführten Resultate wäre es durchaus denkbar, dass eine genauere POS-Unterteilung weiter Aufschluss über einflussreiche Wortarten im Linkskontext liefern würde. Allerdings hat sich in einigen Versuchen, diese vorzunehmen, gezeigt, dass die hier verwendete Stichprobe dafür nicht groß genug war, weshalb ich in diesem Rahmen zunächst nur den Einfluss der drei angegebenen Instanzen untersuchen konnte.

5.2.3 Einfluss der lokalen Variablen

Unter den lokalen Variablen hat sich, wie zu erwarten, die Kleinschreibung von EN für alle Programme als relevant erwiesen. Besteht ein EN ausschließlich aus Minuskeln, erhöht sich die Wahrscheinlichkeit für dessen Fehlklassifikation durchschnittlich beinahe um das Neunfache. Abgesehen von Unternehmensnamen, die bisweilen zum Zwecke der Differenzierung und Hervorhebung ihres Designats durchgehend kleingeschrieben werden (vgl. **nuebling12**), ist diese Fehlerquelle vor allem durch die Beschaffenheit des CatTle.de.12 bedingt. In Forentexten und ähnlichem eher spontansprachlichem Material, woraus ca. 35 % der in der Stichprobe enthaltenen Tokens stammen, wird die orthographische Norm der Substantivgroßschreibung häufig nicht berücksichtigt und es ist stattdessen üblich, ausschließlich Minuskeln zu verwenden. Dadurch wird zusätzlich zur Unterscheidung der EN von den Appellativen auch die Abgrenzung der Onyme zu allen anderen Wörtern erschwert, was zu schlechterer Performanz führt. Dafür, dass es sich bei der Großschreibung um ein für die Algorithmen der Programme sehr wichtiges Merkmal bei der EN-Erkennung handelt, sprechen auch die Odds-Ratios auf Seiten der False Positives. Dass die Werte alle nah bei Null liegen, zeigt, dass ein kleingeschriebenes Wort im Allgemeinen eine sehr geringe Chance hat, fälschlicherweise als EN getaggt zu werden.
Anders verhält es sich mit der ausschließlichen Verwendung von Majuskeln, welche ihrerseits zumindest für den HGC-Klassifizierer die Wahrscheinlichkeit, einen False Positive zu erhalten, signifikant erhöht. Dass durchgehend groß geschriebene False Positives vor allem in der Kategorie *ORG* auftauchen, lässt sich dadurch erklären, dass hauptsächlich Unternehmens- und Produktnamen zu Werbezwecken auf Abweichungen von orthographischen Normen zurückgreifen oder auch Abkürzungen aus Großbuchstaben enthalten (vgl **nuebling12**), während dies bei anderen EN-Klassen eher unüblich ist. Wird nun ein Wort, bei dem es sich nicht um einen EN handelt, in Großbuchstaben abgekürzt oder zur Hervorhebung in Majuskeln gesetzt, scheint es plausibel, dass ein Programm eher zu dessen fälschlicher Klassifizierung als Organisationsname neigt. Dies erklärt somit gleichzeitig einen Teil des Einflusses, den Abkürzungen auf die Fehlerwahrscheinlichkeit haben. Allerdings ist anzumerken, dass auch klein geschriebene Akronyme unter den False Positives zu finden sind.
Das Vorhandensein von Sonderzeichen steigert, ähnlich wie bei den kontextuellen Variablen der Definitartikel und die Interpunktion, für beide abhängigen Variablen die Fehlerwahrscheinlichkeit. Hier fällt eine Interpretation allerdings leichter, da Sonderzeichen in Tokens ein seltenes Ereignis sind und weder für EN noch für Nicht-EN eine Charakteristik darstellen. Somit werden in dem aus Zeitungsdaten bestehenden Trainingsmaterial insgesamt eher wenige solche Tokens vorgekommen sein, die, von Bindestrichen abgesehen, etwas anderes als Buchstaben enthielten, weshalb die Programme kaum genügend statistische Erfahrung sammeln konnten, um mit solchen Tokens umzugehen. Hinzu kommt, dass, wie die explorative Fehlerdurchsicht

bestätigt hat, beispielsweise in Foren gewisse Sonderzeichen kommunikative Funktionen haben und daher in dieser Textsorte häufiger auftreten. Hierzu zählen unter anderem Emoticons aber auch das *@*, welches zur Adressatenkennzeichnung ohne Leerzeichen vor einen Namen gesetzt wird, wie es beispielsweise in *@angela* im untersuchten Teilkorpus vorkommt. Somit kann es in eher spontansprachlichen Texten zur Häufung von Sonderzeichen enthaltenden Tokens kommen und die Fehleranzahl steigen.

Neben den graphematischen haben sich auch einige morphologische und morphosyntaktische Eigenschaften von Wörtern als relevant für die Entstehung, bzw. die Vermeidung von Fehlern erwiesen. So verringert sowohl das Vorhandensein von Derivations- als auch das von Flexionssuffixen die Wahrscheinlichkeit für False Positives für den deWaC-Klassifizierer und den TreeTagger, während bei letzterem zumindest für das Vorhandensein von Derivationssuffixen wiederum eine Steigerung der Wahrscheinlichkeit für False Negatives festzustellen ist. Obgleich beide Suffixkategorien in die gleiche Richtung zu wirken scheinen, erfordert ihr Einfluss möglicherweise unterschiedliche Erklärungsansätze. So ist davon auszugehen, dass besonders der TreeTagger, der nicht nur EN, sondern auch andere Wortarten erkennt, anhand von Lernmaterialien bereits Generalisierungen über viele der für das Deutsche charakteristischen Derivationssuffixe abgeleitet hat. Trifft das Programm nun im Tagging-Prozess auf ein ihm unbekanntes Token, kann das Vorhandensein eines Derivationssuffixes dabei helfen, dessen Wortart richtig zu bestimmen und so einen Fehler zu vermeiden. Auf der anderen Seite kann ein Derivationssuffix bei einem EN dazu führen, dass dieser fälschlicherweise zu jener Wortart gezählt wird, für welche das Suffix charakteristisch ist. Dies legt auch der Odds-Ratio in der rechten Hälfte der Tabelle nahe, der besagt, dass es für Derrivationssuffix tragende EN um das rund Sechsfache wahrscheinlicher ist, vom TreeTagger nicht als solche erkannt zu werden. Ob dieser Erklärungsansatz auch für Flexionssuffixe greift ist fraglich, da diese nicht per se für eine bestimmte Wortart charakteristisch sind. Vielmehr wäre hier ein Zusammenhang mit der bereits in Abschnitt 2.2.2 angemerkten für Onyme typischen Minimalflexion möglich. Da das Paradigma von EN in der Regel nur mehr zwei Wortformen aufweist und sich lediglich der Genitiv Singular und die aufgrund der Monoreferenz eher selten auftretenden Formen der Pluraldeklination durch das Flexiv *-s* von der Nennform unterscheiden, scheint eine Minderung der Wahrscheinlichkeit für die fälschliche Kennzeichnung eines flektierten Appellativs als EN durchaus plausibel.

Die Interpretation der Odds-Ratios der beiden verbleibenden lokalen Variablen *Ambig* und *Fremdsprachlich* gestaltet sich etwas klarer als jene der morphologischen Einflussgrößen. Die semantische Eigenschaft der Ambiguität von EN hat sich den Erwartungen entsprechend in den meisten Fällen als hochgradig signifikant erwiesen. Da es hier um solche Belege geht, die wie die Beispiele für die Kategorien *LOC* und *ORG* aus Abschnitt 4.3 zwar EN sind, jedoch auch außerhalb derselben vorkommen, ist es, wie bereits erwähnt, lediglich für Onyme interessant, diese unabhängige

Variable zu betrachten. Aus diesem Grund ist die betreffende Zeile in der linken Hälfte der Tabelle leer. Es ist offensichtlich, dass allen EN-Klassifizierern das Erkennen von EN, die nicht-onymische Homonyme haben, besonders schwer fällt. Für den HCG-Klassifizierer steigert sich die Fehlerwahrscheinlichkeit sogar beinahe um das Vierzehnfache. Dass für den TreeTagger die Werte dieser Variablen keine Rolle spielen, erklärt sich durch die Beschaffenheit der in Abschnitt 4.2 beschriebenen STTS-Richtlinien, nach denen solchen Tokens stets ihre genuine Wortart und nicht das für Onyme vorgesehene Label *NE* zuzuweisen ist.
Mit einer ähnlich großen Sicherheit war der Einfluss von fremdsprachlichem Material auf die Performanz der Programme vorherzusehen. Handelt es sich bei dem zu klassifizierenden Token um ein Wort aus einer anderen Sprache, steigt die Wahrscheinlichkeit für eine Fehlentscheidung bei einigen Programmen um beachtliche Faktoren. Unklar bleibt jedoch, ob der HGC-Klassifizierer und der SemiNER dazu tendieren, fremdsprachliche Tokens eher den EN zuzuweisen, und der HGC-Klassifizierer im Gegensatz dazu eher zu einer Klassifizierung als Nicht-EN neigt, wie es die Odds-Ratios zunächst suggerieren könnten. Für die Beantwortung dieser Frage wäre es hilfreich, die Einflussrichtung für die jeweils andere Hälfte der Tabelle zu kennen. Da sich diese allerdings in den Modellen für keines der drei Programme als signifikant herausgestellt hat, ist eine solche Schlussfolgerung ohne eine genauere Untersuchung der Daten auf diese Frage hin nicht gerechtfertigt.

5.2.4 Einfluss des Registers

In einigen Fällen erwiesen sich auch die unterschiedlichen Register der Dokumente, aus denen die zu klassifizierenden Tokens stammten, als signifikant Einfluss nehmend. Allerdings ergibt sich hier erneut ein uneinheitliches Bild, bei dem die selbe Ausprägung einer unabhängigen Variable bei den verschiedenen Programmen in entgegengesetzte Richtungen wirkt. Dass für den TreeTagger die Wahrscheinlichkeit für die Entstehung eines False Negatives bei Tokens aus eher spontansprachlichen Texten um die Hälfte sinkt, ist etwas unerwartet, hätte man doch vermutet, dass die orthographischen Eigenschaften dieses Registers die Fehlerwahrscheinlichkeit auf beiden Seiten ansteigen lassen. Besonders verwunderlich ist dieser Wert zudem im Vergleich mit jenen des deWaC- und des HGC-Klassifizierers, aus denen hervorgeht, dass bei diesen beiden Programmen die Wahrscheinlichkeit für False Negatives in der Kategorie *Quasi-Spontaneous*, wie zu erwarten, ansteigt. Ein klar interpretierbares Bild liefern somit lediglich die Werte für die False Positives, deren Entstehungswahrscheinlichkeit durch die eher spontane Sprache in allen signifikanten Fällen erhöht wird. Dies lässt sich leicht durch die orthographischen Besonderheiten wie beispielsweise das Verwenden von Sonderzeichen und Emoticons in Foren und ähnlichen Internetplattformen erklären. Hinzu kommt, dass in spontansprachlichen Internetbeiträgen bisweilen prosodischer Fokus, zum Beispiel durch durchgängige Großschreibung oder Vokalverdopplung, transkribiert wird. Die Ergebnisse unterstützen somit die Annahme, dass Internetkorpora für die EN-Erkennung eine

besondere Herausforderung darstellen.

Tabelle 5: Performanz der Programme nach den Werten der Kategorie Mode

Mode	Precision	Recall	F_1-Measure
TreeTagger			
Written	78,1	66,6	71,9
Quasi-Spontaneous	50,8	62,6	56,1
deWaC			
Written	75,7	50,6	60,7
Quasi-Spontaneous	53,3	32,4	40,3
HGC			
Written	78,4	46,5	58,4
Quasi-Spontaneous	52,9	29,0	37,4
SemiNER			
Written	66,0	46,9	54,8
Quasi-Spontaneous	31,7	27,0	29,2

Untermauert wird diese These auch durch den direkten Vergleich der Performanz der Programme auf dem schriftsprachlichen Teil der Stichprobe mit der auf dem eher spontansprachlichen Teil. Die Tabelle 5 zeigt, dass der TreeTagger einen um 15%, der deWaC- und der HGC-Klassifizierer einen um 20% und der SemiNER sogar einen um 25% schlechteren F_1-Measure auf den spontansprachlichen Daten erreichen als auf jenen Dokumenten, die der Kategorie *Written* angehören.
Lediglich im Falle des deWaC-Klassifizierers wird die Fehlerwahrscheinlichkeit auch von einer Ausprägung der unabhängigen Variable *Audience* beeinflusst. So verringert sich für Tokens aus der Kategorie *Informed*, welche solche Texte umfasst, die sich an einen eingeschränkten, über ein spezifisches Wissen verfügenden Personenkreis richten, die Wahrscheinlichkeit für eine fälschliche Klassifizierung als EN. Hierfür gibt es mehrere Erklärungsmöglichkeiten. Zum einen wäre es denkbar, dass Texte, die an ein begrenztes, eingeweihtes Publikum gerichtet sind, tendenziell eine größere Anzahl spezialisierter Begriffe und somit womöglich auch unbekannter EN enthalten. Dann müsste allerdings auch eine Steigerung der Wahrscheinlichkeit für False Neagtives zu beobachten sein, was nicht der Fall ist. Zum anderen könnte eine denkbare Ursache darin liegen, dass in diesen Dokumenten vergleichsweise weniger EN vorhanden sind als in den anderen Audience-Ausprägungen. Da für eine eindeutige Erklärung jedoch eine genaue linguistische Analyse der betreffenden Dokumente nötig wäre, die im Rahmen dieser Arbeit nicht geleistet werden kann, ist eine Interpretation auf der Grundlage dieses einen Ergebnisses kaum möglich.

5.3 Zusammenfassung der Fehlerklassifikation

Mithilfe der Regressionsmodelle konnten für die vier verschiedenen Programme einige lokale, kontextuelle und die Textsorte betreffende Eigenschaften festgestellt werden, die Einfluss auf die Güte der EN-Erkennung nehmen. Da in den unabhängigen Variablen sowohl solche Charakteristiken repräsentiert sind, welche die Algo-

rithmen der Programme berücksichtigen, als auch die in einer vorangehenden qualitativen Fehlerauswertung aufgestellten Hypothesen überprüft wurden, kann von einer umfassenden Untersuchung der Fehlerursachen gesprochen werden. Dies spiegelt sich auch in den Werten für die Beurteilung der Güte der Regressionsmodelle wider, die in der Tabelle 7 im Anhang zu aufgeführt sind. Wenngleich einige Ergebnisse nicht definitiv interpretiert werden können, haben sich doch viele der im Vorhinein aufgestellten Vermutungen über Fehler begünstigende Faktoren als zutreffend erwiesen. Auf dieser Grundlage möchte ich im nun folgenden Kapitel die Vor- und Nachteile verschiedener Lernmechanismen für die Programme diskutieren und Möglichkeiten besprechen, die Performanz der EN-Erkenner in Zukunft zu verbessern.

6 Möglichkeiten zur Performanzverbesserung

6.1 Potential und Grenzen der verschiedenen Lernmaterialien

Sowohl im CoNLL-2002 Shared Task als auch bei der Wiederholung des Wettbewerbs im Folgejahr lag ein Hauptaugenmerk darauf, über annotierte Trainingsdaten hinaus zusätzliche Materialien wie Namenslisten, sogenannte Gazetteers, oder unannotierte Korpora zur möglichst effizienten Performanzsteigerung einzusetzen (**tjong03**, **tjong02**). Für das Deutsche, welches zusammen mit dem Englischen die Ausgangssprache für den Wettbewerb im Jahr 2003 darstellte, führte dies zu einer Fehlerreduktion von 3% bis 15%. Auch die Entwickler der in dieser Arbeit evaluierten EN-Erkenner wandten, wie bereits in Kapitel 3 erläutert wurde, über die Trainingsdaten hinaus weitere Lernmethoden an. Diese sollen nun im Folgenden noch einmal genauer beleuchtet und auch unter Bezugnahme auf die Fehlerklassifikation auf ihre Möglichkeiten und Grenzen hin untersucht werden.

6.1.1 Clustering in unannotierten Korpora

Alle drei hier besprochenen EN-Erkenner wandten die Technik des Clusterings in unannotierten Korpora an. Hierbei unterschieden sich die Ansätze sowohl in der Größe der Datensätze als auch in der Anzahl der gebildeten Cluster. Während Faruqui und Padó das gesamte, 175 Mio. Tokens umfassende HGC und die gleiche Tokenanzahl des deWaC für diesen maschinellen Lernschritt nutzten, enthielt das von Chrupała & Klakow verwendete ECI-Korpus nur mehr ca. 34 Mio. Tokens. Wenngleich es sich mit dem Umfang des annotierten Trainingsmaterials verglichen in allen Fällen um beachtliche Datenmengen handelt, ist doch ein bemerkenswerter Größenunterschied zwischen den Datensätzen, die Faruqui und Padó verwendeten und jenem, der als Clusteringgrundlage für den SemiNER diente, festzustellen. Was die Clusteranzahl betrifft, wählten die Entwickler des SemiNER allerdings mit 1000 Clustergruppen eine feinere Unterteilung als Faruqui und Padó, die für den HGC-Klassifizierer mit 600 Clustergruppen das beste Ergebnis erzielten und für den des

deWaC mit 400. Zwar erzielten die Programme auf den CoNLL-2003-Testdaten zunächst eine vergleichbare Verbesserung durch den Clustering-Schritt, auf jenen des CatTle.de.12 fiel der Performanzabfall jedoch unterschiedlich stark aus. So büßte der SemiNER ganze 27,5% F_1-Measure ein, während sich der deWaC- und der HGC-Klassifizierer mit 21,2%, bzw. 25,6% als robuster erwiesen. Dies könnte darauf hindeuten, dass bei der Programmentwicklung ein größeres Korpus und weniger Cluster einem kleineren Korpus mit einer genaueren Einteilung vorzuziehen ist. Es wäre sicherlich interessant, hier weitere Tests mit größeren Cluster-Datensätzen durchzuführen, um diese Hypothese zu überprüfen.

6.1.2 Die berücksichtigten Merkmale

Die von den Algorithmen bei der EN-Erkennung berücksichtigten kontextuellen und lokalen Merkmale stehen in einer direkten Verbindung zu den in Abschnitt 5.2 besprochenen Fehlerquellen, welche mit Hilfe der Regressionsmodelle identifiziert wurden. Auf der einen Seite sind die Generalisierungen über typische syntaktische, morphologische, semantische und graphematische Eigenschaften von EN ein grundlegendes Hilfsmittel für deren korrekte Klassifizierung. Auf der anderen Seite können sie, wie die Fehlerklassifikation gezeigt hat, auch in die Irre leiten und zu Falschkategorisierungen führen, da es sich bei ihnen nie um hinreichende Merkmale, sondern lediglich um prototypische Eigenschaften handelt. Zwar berücksichtigen der Stanford NER, der den beiden Klassifizierern von Faruqui und Padó zugrunde liegt, und der SemiNER im Wesentlichen die gleichen Eigenschaften bei der Klassifizierung, jedoch bezieht der EN-Erkenner aus Stanford einige wenige kontextuelle Merkmale mehr in die Entscheidungen mit ein. So analysiert er bei den das zu klassifizierende Token umgebenden Wörtern auch die POS-Tags, während der SemiNER ausschließlich die Wortform der adjazenten Tokens beachtet. Zusätzlich wird noch für vier Tokens im Links- und Rechtskontext überprüft, ob es sich bei ihnen um Wörter oder aber andere Zeichen wie Interpunktion oder Satzgrenzen handelt (**finkel05**). Wenngleich der Tabelle 4 zu entnehmen ist, dass beispielsweise der Einfluss von EN im Kontext auf die Fehlerwahrscheinlichkeit beim SemiNER besonders groß ist, lässt sich dennoch vor allem im Vergleich zu den ebenfalls hohen Werten der Odds-Ratios des deWaC-Klassifizierers ohne weitere Untersuchungen nicht sagen, ob diese etwas reduzierte Merkmalsrepräsentation den SemiNER fehleranfälliger macht und zu dessen größerem Performanzverlust auf den CatTle.de.12-Daten beiträgt. Vielmehr ist möglicherweise auch hier ein Rückbezug auf den Umfang der Clustering-Daten angemessen. So scheint die Annahme plausibel, dass die von den Programmen abgeleiteten Regeln über die Merkmale von EN mit einer größeren Menge an gesehenem Textmaterial der tatsächlichen sprachlichen Realität immer näher kommen und somit weniger Fehler entstehen. In Anbetracht des Umstands, dass die Qualität der Klassifizierer nicht wesentlich von der Textsorte der Datengrundlage für das Clustering abzuhängen scheint und sich der deWaC-Klassifizierer sogar als robuster als der HGC-Klassifizierer erwies, stellen Internetkorpora aufgrund ihres Umfangs ei-

ne interessante Option für diesen Lernschritt dar. Die Merkmalsrepräsentation der Algorithmen noch zu erweitern und beispielsweise die in Abschnitt 2.2.1 und 2.2.4 besprochenen besonderen Vokal- und Konsonantenkombinationen, die meist den EN vorbehalten sind, mit einzubeziehen scheint daher nicht allzu vielversprechend und würde vermutlich lediglich zu einer längeren Laufzeit führen, ohne die Ergebnisse der Programme wesentlich zu beeinflussen.

6.1.3 Namenslisten

Mit den Infobox-Einträgen aus Wikipedia nutzten Chrupała & Klakow eine weitere Ressource zur Verbesserung der Performanz. Allerdings zeigte diese einen deutlich niedrigeren Effekt auf die Güte des SemiNER als die Clustering-Daten. Während die Wikipedia-Informationen den F_1-Measure um kaum mehr als 5% anhoben, führten die Clustering-Daten im Vergleich zur ausschließlichen Verwendung der annotierten Trainingsdaten zu einer Verbesserung von ca. 15%. Zudem ergänzten sich die in den beiden Lernansätzen gewonnenen Informationen offenbar nicht, was daraus ersichtlich wird, dass die Infoboxen zusätzlich zu den Cluster-Daten kaum mehr eine Performanzsteigerung bewirkten. Die Entwickler folgerten daraus, dass die den Wikipedia-Artikeln entnommenen Informationen kaum wesentlich Neues zu jenen aus den Clustering-Daten beitragen (**chrupala10**). Insofern, als die Wikipedia-Informationen nicht in erster Linie als Namensliste, sondern vorrangig zum Ableiten der Korrelation zwischen bestimmten EN-Kategorien und den assoziierten Infobox-Labels genutzt wurde, ist aufgrund dieser Erkenntnis die Verwendung von Gazetteers jedoch keinesfalls als ineffektiv zu bewerten. Gegen eine solche Schlussfolgerung sprächen auch die Ergebnisse des CoNLL-2003 Shared Tasks, bei dem sich laut **tjong03** für das Deutsche gerade kein merklicher Unterschied zwischen der Performanzverbesserung durch die Verwendung von unannotierten Daten und der durch den Einsatz von Namenslisten habe feststellen lassen. Selbstverständlich können Gazetteers aufgrund des Umfangs und der Unabgeschlossenheit der Menge von EN stets nur einen Teil der in einem Korpus vorkommenden Onyme abdecken, jedoch befanden sich unter den gesichteten False Negatives des öfteren auch äußerst prominente Namen wie *Obama* oder *Schalke*, die im DECOW2012 immerhin eine Frequenz von 7,8 pMW und 10,5 pMW aufweisen. Gerade solche Fehler könnten eventuell mit der richtigen Auswahl von aktualisierten Namenslisten vermieden werden. Mittlerweile gibt es einige sehr interessante Ansätze zur Erstellung solcher Datenbanken. Eine ähnliche Vorgehensweise wie jene von Chrupała und Klakow wird zum Beispiel im Projekt DBpedia (**mendes12**) verwendet, bei welchem unter anderem auch für das Deutsche Informationen aus Wikipedia extrahiert und aufbereitet werden. Allerdings beschränken sich die Entwickler der Datenbank nicht auf Infoboxen, sondern beziehen auch Informationen aus Artikelüberschriften und den meist dem eigentlichen Wikipedia-Artikel vorangehenden Abstracts mit ein. Zur Zeit umfasst der deutsche Datensatz rund 1,25 Mio. Einträge[6], bei denen

[6] http://wiki.dbpedia.org/Datasets/DatasetStatistics#h251-3, letzter Zugriff am 09.06.2013.

es sich jedoch nicht ausschließlich um EN handelt. Nichtsdestotrotz übersteigt die Anzahl die 218 000 Artikel, zu denen Chrupała und Klakow Informationen aus Wikipedia extrahierten, bei weitem, sodass es durchaus denkbar ist, dass eine größere Menge an EN in DBpedia enthalten sein könnte. Eine ausschließlich auf EN ausgerichtete Datenbank, deren Informationen ebenfalls Wikipedia entnommen sind, ist mit HeiNER (**wentland08**) gegeben. Da diese Ressource speziell für EN-Erkennung gedacht und aufbereitet ist, bringt sie einige interessante Zusatzinformationen mit sich. So werden zunächst die EN aus der englischen Wikipedia-Version in alle 253 in der Online-Enzyklopädie verfügbaren Sprachen übersetzt. Für 16 Sprachen, unter anderem für das Deutsche, stellen die Entwickler darüber hinaus Disambiguierungs-Wörterbücher zur Verfügung, in denen alle in Wikipedia aufgeführten onymischen Denotate eines bestimmten EN enthalten sind. Um die korrekte Klassifizierung solcher polysemer EN zu gewährleisten, extrahierten **wentland08** für alle Instanzen im Disambiguierungs-Wörterbuch den Links- und Rechtskontext der auf diese Einträge verweisenden Hyperlinks aus anderen Wikipedia-Seiten. Mit einem Umfang von mehr als 1,5 Mio. disambiguierten EN[7] stellt HeiNER eine äußerst vielversprechende Ressource für die EN-Erkennung dar.
Zusätzlich zu diesen etwas aufwändiger aufbereiteten Datenbanken existieren andere frei zugängliche Gazetteers wie beispielsweise aus Telefonbüchern extrahierte Personennamen oder Listen mit Ortsnamen, die auch komplexe EN enthalten. In Anbetracht dieser Fülle an verfügbaren Ressourcen wäre es durchaus interessant, zu überprüfen, ob und wie sehr die Verwendung dieser Namenslisten die Performanz der Programme verbessern könnte.

6.2 Richtlinien

Eine weitere Möglichkeit, auf die Güte der EN-Erkennung Einfluss zu nehmen, stellen die Richtlinien insofern dar, als sie den Maßstab festlegen, anhand dessen die Programme evaluiert werden. Bei der Erstellung derselben spielen meist nicht nur die möglichst onomastisch korrekte Einteilung, sondern mehrere Faktoren eine Rolle. Dass EN-Erkennung nicht nur für die theoretische Linguistik von Interesse ist, sondern auch für viele automatische Sprachverarbeitungsschritte wie beispielsweise Informationsextraktion einen notwendigen Vorverarbeitungsschritt darstellt, kann wie im Fall der Annotationsrichtlinien des CoNLL-2003 Shared Tasks zu einigen linguistisch kaum motivierbaren EN-Definitionen führen. So werden unter anderem auch nominale und adjektivische Derivate von Onymen wie beispielsweise der bereits erwähnte *Singapur-Besucher* oder *französisch* und *Marx'sch* in die Kategorie *MISC* eingeordnet, obwohl sie keine der funktionalen Eigenschaften von EN aufweisen. Eventuell ist dies für die maschinelle Weiterverarbeitung der Texte von Interesse, jedoch ergibt sich dadurch weder eine linguistisch korrekte Annotation, noch ist es eine leichte Aufgabe für die Programme, solche Derivate zu erkennen, was auch die

[7] http://heiner.cl.uni-heidelberg.de/, letzter Zugriff am 09.06.2013.

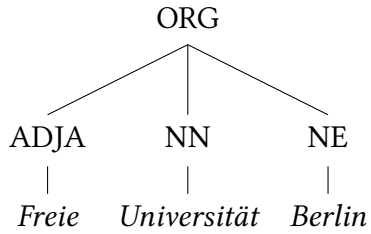
besonders schlechte Performanz der EN-Erkenner in der Kategorie *MISC* nahelegt. Bei einer im Jahre 2006 vorgenommenen Überarbeitung der CoNLL-Richtlinien wurde diesem Umstand genüge getan und explizit angegeben, dass EN-Derivate nicht mehr gekennzeichnet werden.
Selbstverständlich ist es durchaus sinnvoll, bei der Erstellung von Richtlinien auch die Grenzen automatischer Wortklassifizierer zu berücksichtigen, da diese sonst unter Umständen Ergebnisse versprechen, die aus technischen Gründen nicht erfüllt werden können. Viele der in dieser Arbeit aufgeführten Probleme, die sich den EN-Erkennern in den verschiedenen Kategorien stellen, werden beispielsweise in den STTS-Richtlinien insofern umgangen, als von vornherein festgelegt wird, dass transparente, also appellativische oder adjektivische Teile eines Onyms entsprechend ihrer genuinen Wortart annotiert werden. Allerdings werden so, wie bereits erläutert, viele nach onomastischen Kriterien durchaus als EN anzusehende Wörter nicht als solche markiert und würden daher bei einer Suchanfrage nicht gefunden. Aus der Sicht der Anwender ist dies also eine ebenso wenig zufriedenstellende Lösung wie Richtlinien, die eine umfangreiche und detaillierte EN-Kennzeichnung versprechen, welche die Programme jedoch nicht durchführen können. In diesem Zusammenhang scheint es angebracht, die Kategorisierung der EN in weitere Unterklassen zu hinterfragen. Da beispielsweise beim HGC-Klassifizierer 75,5% der False Positives und 26% der False Neagatives zwar richtig als EN erkannt, jedoch einer falschen Unterkategorie zugewiesen wurden, wäre unter Umständen gemeinsam mit Onomastikern und Programmentwicklern zu diskutieren, inwiefern diese Unterteilung sinnvoll, bzw. notwendig ist. So zeigt die Tabelle 6, dass die Performanz der EN-Erkenner, wenn man die zwar erkannten, aber mit dem falschen Kategorie-Label versehenen Tokens nicht als Fehler wertet, deutlich ansteigt. In Klammern sind die erreichten Ergebnisse bei Berücksichtigung der Unterklassen aus Tabelle 3 zum Vergleich angegeben. Da der TreeTagger von vornherein keine weitere Unterteilung vornimmt, ist hier selbstverständlich keine Veränderung festzustellen.

Tabelle 6: Performanz der Programme ohne Berücksichtigung der Einteilung in EN-Unterklassen. In Klammern ist die Performanz unter Berücksichtigung der Einteilung zum Vergleich angegeben

	Precision	**Recall**	**F_1-Measure**
TreeTagger	**69,5**	**66,1**	**67,8**
deWaC	**83,2** (69,8)	**54,2** (45,5)	**65,7** (55,2)
HGC	**84,7** (71,6)	**49,2** (41,6)	**62,2** (52,6)
SemiNER	**72,8** (55,1)	**54,6** (41,3)	**62,4** (47,2)

Evaluiert man die Programme also lediglich in Bezug darauf, ob sie richtig zwischen EN und Nicht-EN unterscheiden, ergibt sich in allen drei Fällen ein um mindestens 10% besserer F_1-Measure. Obgleich allen voran die Precision steigt, verbessert sich auch der Recall um zwsichen 7,6% und 13,3%. Zwar liegen diese Werte noch immer unter jenen des TreeTaggers, bedenkt man aber die unterschiedliche Anzahl der EN,

von denen die Goldstandards aufgrund der jeweiligen Richtlinien ausgehen, kennzeichnen sowohl der SemiNER als auch die beiden Klassifizierer von Faruqui und Padó in absoluten Zahlen mehr EN als der TreeTagger, was sie für die Aufbereitung von Korpora durchaus interessant macht. Selbstverständlich ist für Fragestellungen, die sich explizit nur auf eine der vier EN-Unterklassen beziehen, ein Korpus, in dem die EN bereits derart sortiert sind, wesentlich attraktiver. Da sich jedoch gezeigt hat, dass vor allem die Kategorien *LOC*, *ORG* und *MISC* teilweise selbst für menschliche Annotatoren schwer voneinander abzugrenzen sind und sich dies auch darin widerspiegelt, dass die Einteilung, wie die Evaluation gezeigt hat, von den Programmen alles andere als zufriedenstellend durchgeführt wird, wäre es womöglich auf dem aktuellen Stand realistischer, diese im Korpus gar nicht erst vornehmen zu wollen.
Wünschenswert, jedoch vermutlich schwer umsetzbar, wäre über die Einteilung der EN in Subklassen hinaus eine automatische Mehrebenenannotation der Onym-Instanzen, wie sie beispielsweise in den TüBa-D/Z-Richtlinien[8] vorgeschlagen wird. Dies scheint vor allem für komplexe EN äußerst sinnvoll, da auf der untersten Ebene zunächst die Wortart nach STTS zugewiesen wird und auf einer höheren Ebene dann auch solche Wörter als dem EN zugehörig markiert sind, die nach STTS nicht das Label *NE* erhalten würden. So entstünde für den Institutionsnamen *Freie Universität Berlin* in etwas vereinfachter Form folgende Repräsentation:



Dieses Annotationsschema wäre für die strukturelle Untersuchung komplexer EN äußerst vorteilhaft und könnte das Erfassen ihrer typischen Bildungsmustern vereinfachen. Da die Evaluation jedoch ergeben hat, dass bisher weder die erste noch die zweite Ebene im Einzelnen zuverlässig annotiert wird, bleibt diese Vorgehensweise wohl fürs Erste der manuellen Annotation vorbehalten.

7 Zusammenfassung und Ausblick

Die im Rahmen dieser Arbeit durchgeführte Evaluation der Programme auf dem Web-Korpus CatTle.de.12 verdeutlichte, dass die automatische EN-Erkennung für das Deutsche bisher keinesfalls zufriedenstellende Ergebnisse liefert. Hatten die Programme bereits auf deutschsprachigen Zeitungsdaten im Vergleich zum Englischen schlechtere Performanz gezeigt, fiel der F_1-Measure auf den Internetdaten noch einmal drastisch ab. Tatsächlich wurde nicht einmal die Hälfte der in der Stichprobe enthaltenen EN richtig erkannt, was den Nutzen der EN-Erkenner vor allem für

[8] http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1201.pdf, letzter Zugriff am 10.06.2013.

solche Korpora, die nicht aus Zeitungstexten bestehen, in Frage stellt. Die Gründe hierfür wurden auf die Eigenschaft von Web-Korpora zurückgeführt, sowohl zu einem beachtlichen Teil aus eher spontansprachlichem Material zu bestehen, welches im Allgemeinen orthographisch weniger normiert und selten editiert ist, als auch eine Vielzahl unterschiedlicher Domänen abzudecken. In Bezug auf die Möglichkeiten zur Verbesserung der Performanz der EN-Erkenner hat sich bei der Fehlerklassifikation zudem herausgestellt, dass, wenngleich kontextuelle, morphologische, morphosyntaktische und semantische Merkmale eine unverzichtbare Basis für die Klassifizierung von Wörtern darstellen, in einigen Fällen auch Fehlkategorisierungen auf sie zurückzuführen sind. Eine detailliertere Merkmalsrepräsentation in den Algorithmen würde daher vermutlich kaum zu besseren Ergebnissen führen. Aufgrund der hohen Kosten scheint auch die Erstellung von größeren, mit EN annotierten Datenmengen keine realistische Möglichkeit zur Performanzsteigerung darzustellen. Zwar könnte es interessant sein, die Programme einmal auf dem im Rahmen dieser Arbeit erstellten Goldstandard zu trainieren und zu überprüfen, welche Ergebnisse die EN-Erkenner auf der Grundlage dieser Lerndaten erzielen. Langfristig kann es jedoch keine Lösung darstellen, zunächst einen Goldstandard aus den eigenen Korpusdaten für das Training der Programme erstellen zu müssen, um einigermaßen zufriedenstellende Ergebnisse zu erhalten. Folglich sind die Entwickler auf die Verwendung unannotierter Daten und Gazetteers angewiesen, die, wie sich bereits im CoNLL-2003 Shared Task gezeigt hat, die Güte der EN-Erkennung durchaus verbessern können. Einen Ansatz zur Minimierung der Fehlerwahrscheinlichkeit könnte die Verwendung möglichst umfangreicher unannotierter Korpora darstellen, deutete doch der Vergleich des SemiNER mit den beiden Klassifizierern von Faruqui und Padó an, dass eine größere Datenmenge einer höheren Anzahl an Clustern vorzuziehen ist. Da es sich hierbei jedoch lediglich um eine Vermutung handelt, die in dieser Arbeit nicht weiter untersucht werden konnte, wäre dies zunächst in einem weiteren Experiment zu überprüfen.

Mit den frei zugänglichen Gazetteers allerdings ist eine zusätzliche, bisher noch wenig erprobte Ressource verfügbar, die sowohl ambige und komplexe EN korrekt zu klassifizieren helfen, als auch den Programmen eine umfangreiche Grundmenge an bekannten EN-Instanzen zur Verfügung stellen könnte. Vor allem Fehler bei hochfrequenten EN könnten so unter Umständen vermieden werden. Somit wäre es durchaus interessant, die Programme unter der Verwendung eines der in Abschnitt 6.1.3 vorgestellten EN-Lexika erneut zu testen und zu überprüfen, ob die Performanz signifikant ansteigt.

Eine wichtige Erkenntnis, die aus der durchgeführten Untersuchung hervorgeht, ist, dass die Unterteilung der EN in semantische Subklassen zusätzlich einen starken Abfall der Performanz der Programme verursacht. Wenngleich eine Vorkategorisierung für solche Fragestellungen, die sich beispielsweise nur auf Orts- oder Personennamen beziehen, durchaus vorteilhaft ist, muss doch bedacht werden, dass auf dem aktuellen Stand eine beachtliche Anzahl derselben zwar als EN erkannt, jedoch in

eine der anderen drei EN-Subklassen eingeteilt würde und somit nicht als Beleg in die Korpusstudie mit einginge. In Anbetracht dieses Umstandes scheint es durchaus sinnvoll, den Verzicht auf diese Unterkategorisierung in Erwägung zu ziehen oder zumindest ausschließlich Personennamen, deren Erkennung sich als am robustesten herausgestellt hat und die auch in der onomastischen Theorie die am wenigsten umstrittene Kategorie repräsentieren dürften, von den anderen EN zu unterscheiden. Zwar müsste dann auf jede Suchanfrage ein zweiter Arbeitsschritt folgen, um die für die Forschungsfrage relevanten Onyme aus der Gesamtmenge heraus zu sortieren, möchte man aber eine repräsentative Korpusstudie durchführen, kann man sich mit den Klassifizierungsergebnissen der EN-Erkenner auf dem jetzigen Stand kaum zufrieden geben.

Abschließend bleibt festzuhalten, dass die automatische EN-Erkennung vor allem auf weniger standartisierten Daten, wie Web-Korpora sie darstellen, für das Deutsche bisher nicht ausreichend verlässlich funktioniert, um den verschiedenen Fragestellungen der Onomastik gerecht zu werden. In dieser Arbeit wurden jedoch einige vielversprechende Ansätze zur Verbesserung der aktuellen Situation aufgezeigt, denen in der weiteren Forschung zur EN-Erkennung und bei der Entwicklung von Software für diese Aufgabe nachgegangen werden könnte.

Software

deWaC- und HGC-Klassifizierer
http://www.nlpado.de/~sebastian/software/ner_german.shtml (letzter Zugriff am 13.07.2013).

SemiNER
http://grzegorz.chrupala.me/seminer.html (letzter Zugriff am 13.07.2013).

Stanford NER
http://www-nlp.stanford.edu/software/CRF-NER.shtml (letzter Zugriff am 13.07.2013).

TreeTagger
http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/ (letzter Zugriff am 13.07.2013).

Richtlinien

CoNLL-2003-Richtlinien
http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt (letzter Zugriff am 13.07.2013).

STTS-Richtlien
http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (letzter Zugriff am 13.07.2013).

TüBa-D/Z-Richtlinien
http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1201.pdf (letzter Zugriff am 10.06.2013).

CatTle.de.12-Kategorisierungsrichtlinien
http://hpsg.fu-berlin.de/cow/files/cowcat2013.pdf (letzter Zugriff am 13.07.2013)

Datenbanken

HeiNER
http://heiner.cl.uni-heidelberg.de/ (letzter Zugriff am 09.06.2013).

DBpedia
http://wiki.dbpedia.org/Datasets/DatasetStatistics#h251-3 (letzter Zugriff am 09.06.2013).

Anhang

Die Tabelle 7 zeigt für die beiden abhängigen Variablen eines jeden Programms die signifikanten Einflussgrößen mitsamt der geschätzten Einflussstärke. Aus den Logit-Koeffizienten kann man hierbei lediglich ablesen, ob der Wert einer unabhängigen Variablen die Wahrscheinlichkeit für eine Fehlklassifizierung erhöht oder verringert, nicht jedoch, um welchen Faktor. Diese Information liefern die Odds-Ratios, die mit ihrem 95%-Konfidenzintervall aufgeführt sind. Der p-Wert zeigt an, als wie signifikant sich der Einfluss der jeweiligen Größe erwiesen hat. Unter Modellgüte sind zudem die Werte zur Beurteilung des Gesamtfits der acht verschiedenen Regressionsansätze angegeben.

Tabelle 7: Ergebnisse der Regressionsanalyse und Modellgüte

TreeTagger								
	False Positives				False Negatives			
Einflussgröße	Logit	Odds	95% CI	p-Wert	Logit	Odds	95% CI	p-Wert
Kontexteigenschaften								
EN an Position -1	2,2	4,4	1,2 – 17,5	*	-2,3	0,5	0,3 – 0,9	*
EN an Position +1	–	–	–	–	-3,0	0,3	0,1 – 0,6	**
Def. Artikel an Position -1	–	–	–	–	3,2	4,7	1,9 – 12,9	**
Lokale Eigenschaften								
Abkürzung	4,3	20,4	5,7 – 98,6	***	–	–	–	–
Sonderzeichen	4,5	24,1	6,7 – 113,6	***	3,1	6,5	2,1 – 24,8	**
Kleinschreibung	-8,2	0,1	0,03 – 0,1	***	3,8	7,9	2,9 – 25,6	***
Flexionssuffix	-3,9	0,1	0,03 – 0,3	***	–	–	–	–
Derivationssuffix	-2,1	0,3	0,1 – 0,9	*	4,3	6,3	2,8 – 15,7	***
Mode: Quasi-Spontaneous	2	2	1,0 – 3,8	*	-2,8	0,5	0,3 – 0,8	**
Modellgüte								
Nagelkerke-R^2	0,62				0,27			
Kreuzvalidierung	0,86				0,63			
deWaC								
	False Positives				False Negatives			
Einflussgröße	Logit	Odds	95% CI	p-Wert	Logit	Odds	95% CI	p-Wert
Kontexteigenschaften								
EN an Position -1	6,0	127,0	32,0 – 888,0	***	-4,5	0,2	0,1 – 0,4	***
EN an Position +1	4,6	23,1	7,1 – 108,8	***	-3,8	0,3	0,1 – 0,5	***
Def. Artikel an Position -1	2,9	4,4	1,7 – 12,8	**	2,9	2,6	1,4 – 5,2	**
Lokale Eigenschaften								
Sonderzeichen	4,1	12,4	4,1 – 46,4	***	3,4	4,7	2,0 – 11,9	***
Kleinschreibung	-8,4	0,001	10^{-4} – 0,003	***	5,0	9,8	4,3 – 25,8	***
Flexionssuffix	-2,5	0,3	0,1 – 0,8	*	–	–	–	–
Ambig	–	–	–	–	4,3	4,8	2,4 – 10,2	***
Fremdsprachliches Material	–	–	–	–	3,8	8,4	2,9 – 27,5	***
Mode: Quasi-Spontaneous	2,1	2,6	1,1 – 6,6	*	2,3	2,0	1,1 – 3,6	*
Audience: Informed	-2,4	0,3	0,1 – 0,8	*	–	–	–	–
Modellgüte								
Nagelkerke-R^2	0,85				0,42			
Kreuzvalidierung	0,89				0,72			

HGC								
	False Positives				False Negatives			
Einflussgröße	Logit	Odds	95% CI	p-Wert	Logit	Odds	95% CI	p-Wert
Kontexteigenschaften								
EN an Position -1	4,8	23,0	7,2 – 93,6	∗ ∗ ∗	-3,1	0,4	0,2 – 0,7	∗∗
EN an Position +1	4,8	24,8	7,4 – 106,1	∗ ∗ ∗	-2,1	0,6	0,3 – 1,0	∗
Lokale Eigenschaften								
Sonderzeichen	–	–	–	–	2,2	3,8	1,3 – 14,4	∗
Kleinschreibung	-7,8	0,02	0,004 – 0,03	∗ ∗ ∗	3,8	5,5	2,4 – 14,6	∗ ∗ ∗
Großschreibung	2,9	20,4	4,1 – 370,8	∗∗	–	–	–	–
Ambig	–	–	–	–	5,0	13,7	5,4 – 42,7	∗ ∗ ∗
Fremdsprachliches Material	2,6	33,7	4,1 – 911,0	∗∗	–	–	–	–
Mode: Quasi-Spontaneous	–	–	–	–	2,3	1,8	1,1 – 3,0	∗
Modellgüte								
Nagelkerke-R^2	0,71				0,27			
Kreuzvalidierung	0,79				0,69			
SemiNER								
	False Positives				False Negatives			
Einflussgröße	Logit	Odds	95% CI	p-Wert	Logit	Odds	95% CI	p-Wert
Kontexteigenschaften								
EN an Position -1	5,0	199,4	38,4 – 3689,0	∗ ∗ ∗	-5,0	0,1	0,05 – 0,3	∗ ∗ ∗
EN an Position +1	6,4	28,5	10,9 – 87,1	∗ ∗ ∗	-6,0	0,03	0,01 – 0,1	∗ ∗ ∗
Def. Artikel an Position -1	–	–	–	–	2,7	2,9	1,4 – 6,5	∗∗
∅ an Position -1	4,5	5,9	2,7 – 13,0	∗ ∗ ∗	2,3	2,4	1,2 – 5,2	∗
Lokale Eigenschaften								
Abkürzung	4,7	146,6	27,9 – 2722,9	∗ ∗ ∗	–	–	–	–
Kleinschreibung	–	–	–	–	4,6	12,7	4,7 – 41,9	∗ ∗ ∗
Ambig	–	–	–	–	4,4	6,8	3,0 – 16,8	∗ ∗ ∗
Fremdsprachliches Material	3,5	48,7	8,0 – 949,1	∗ ∗ ∗	–	–	–	–
Mode: Quasi-Spontaneous	4,5	4,6	2,4 – 9,2	∗ ∗ ∗	–	–	–	–
Modellgüte								
Nagelkerke-R^2	0,72				0,49			
Kreuzvalidierung	0,90				0,71			